

Olivier Colliot *Editor*

Machine Learning for Brain Disorders

OPEN ACCESS



Humana Press

NEUROMETHODS

Series Editor
Wolfgang Walz
University of Saskatchewan
Saskatoon, SK, Canada

For further volumes:
<http://www.springer.com/series/7657>

Neuromethods publishes cutting-edge methods and protocols in all areas of neuroscience as well as translational neurological and mental research. Each volume in the series offers tested laboratory protocols, step-by-step methods for reproducible lab experiments and addresses methodological controversies and pitfalls in order to aid neuroscientists in experimentation. *Neuromethods* focuses on traditional and emerging topics with wide-ranging implications to brain function, such as electrophysiology, neuroimaging, behavioral analysis, genomics, neurodegeneration, translational research and clinical trials. *Neuromethods* provides investigators and trainees with highly useful compendiums of key strategies and approaches for successful research in animal and human brain function including translational “bench to bedside” approaches to mental and neurological diseases.

Machine Learning for Brain Disorders

Edited by

Olivier Colliot

CNRS, Paris, France

Editor

Olivier Colliot
CNRS
Paris, France

ISSN 0893-2336

Neuromethods

ISBN 978-1-0716-3194-2

<https://doi.org/10.1007/978-1-0716-3195-9>

ISSN 1940-6045 (electronic)

ISBN 978-1-0716-3195-9 (eBook)

© The Editor(s) (if applicable) and The Author(s) 2023

Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Humana imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.

The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

Preface to the Series

Experimental life sciences have two basic foundations: concepts and tools. The Neuromethods series focuses on the tools and techniques unique to the investigation of the nervous system and excitable cells. It will not, however, shortchange the concept side of things as care has been taken to integrate these tools within the context of the concepts and questions under investigation. In this way, the series is unique in that it not only collects protocols but also includes theoretical background information and critiques which led to the methods and their development. Thus it gives the reader a better understanding of the origin of the techniques and their potential future development. The *Neuromethods* publishing program strikes a balance between recent and exciting developments like those concerning new animal models of disease, imaging, in vivo methods, and more established techniques, including, for example, immunocytochemistry and electrophysiological technologies. New trainees in neurosciences still need a sound footing in these older methods in order to apply a critical approach to their results.

Under the guidance of its founders, Alan Boulton and Glen Baker, the Neuromethods series has been a success since its first volume published through Humana Press in 1985. The series continues to flourish through many changes over the years. It is now published under the umbrella of Springer Protocols. While methods involving brain research have changed a lot since the series started, the publishing environment and technology have changed even more radically. Neuromethods has the distinct layout and style of the Springer Protocols program, designed specifically for readability and ease of reference in a laboratory setting.

The careful application of methods is potentially the most important step in the process of scientific inquiry. In the past, new methodologies led the way in developing new disciplines in the biological and medical sciences. For example, Physiology emerged out of Anatomy in the nineteenth century by harnessing new methods based on the newly discovered phenomenon of electricity. Nowadays, the relationships between disciplines and methods are more complex. Methods are now widely shared between disciplines and research areas. New developments in electronic publishing make it possible for scientists that encounter new methods to quickly find sources of information electronically. The design of individual volumes and chapters in this series takes this new access technology into account. Springer Protocols makes it possible to download single protocols separately. In addition, Springer makes its print-on-demand technology available globally. A print copy can therefore be acquired quickly and for a competitive price anywhere in the world.

Saskatoon, SK, Canada

Wolfgang Walz

Preface

Machine learning (ML) is at the core of the tremendous progress in artificial intelligence in the past decade. ML offers exciting promises for medicine. In particular, research on ML for brain disorders is a very active field. Neurological and psychiatric disorders are particularly complex and can be characterized using various types of data. ML has the potential to exploit such rich and complex data for a wide range of benefits including a better understanding of disorders, the discovery of new biomarkers, assisting diagnosis, providing prognostic information, predicting response to treatment and building more effective clinical trials.

Machine learning for brain disorders is an interdisciplinary field, involving concepts from different disciplines such as mathematics, statistics and computer science on the one hand and neurology, psychiatry, neuroscience, pathology and medical imaging on the other hand. It is thus difficult to apprehend for students and researchers who are new to this area. The aim of this book is to provide an up-to-date and comprehensive guide to both methodological and applicative aspects of ML for brain disorders. This book aims to be useful to students and researchers with various backgrounds: engineers, computer scientists, neurologists, psychiatrists, radiologists, neuroscientists, etc.

Part I presents the fundamentals of ML. The book starts with a non-technical introduction to the main concepts underlying ML (Chapter 1). The main classic ML techniques are then presented in Chapter 2. Even though not recent for most of them, these techniques are still useful for various tasks. Chapters 3–6 are devoted to deep learning, a family of techniques which have achieved impressive results in the past decade. Chapter 3 describes the basics of deep learning, starting with simple artificial neural networks and then covering convolutional neural networks (CNN) which are a standard family of approaches that are mainly (but not only) used for imaging data. Those architectures are feed-forward, meaning that information flows only in one direction. On the contrary, recurrent neural networks (RNN), presented in Chapter 4, involve loops. They are particularly adapted to sequential data, including longitudinal data (repeated measurements over time), time series and text. Chapter 5 is dedicated to generative models: models that can generate new data. A large part is devoted to generative adversarial networks (GANs), but other approaches such as diffusion models are also described. Finally, Chapter 6 presents transformers, a recent approach which is now the state-of-the-art for natural language processing and has achieved impressive results for other applications such as imaging.

Part II is devoted to the main types of data used to characterize brain disorders. These include clinical assessments (Chapter 7), neuroimaging (including magnetic resonance imaging—MRI, positron emission tomography—PET, computed tomography—CT, single-photon emission computed tomography—SPECT, Chapter 8), electro- and magnetoencephalography (EEG/MEG, Chapter 9), genetic and omics data (including genotyping, transcriptomics, proteomics, metabolomics, Chapter 10), electronic health records (EHR, Chapter 11), mobile devices, connected objects and sensor data (Chapter 12). The emphasis is put on practical aspects rather than an in-depth description of the underlying data acquisition techniques (which can be complex, for instance in the case of neuroimaging or omics data). The corresponding chapters describe which information do these data provide,

how they should be handled and processed and which features can be extracted from such data.

Part III covers the core methodologies of ML for brain disorders. Each chapter is devoted to a specific medical task that can be addressed with ML, presenting the main state-of-the-art techniques. Chapter 13 deals with image segmentation, a crucial task for extracting information from images. Image segmentation techniques allow delineating anatomical structures and lesions (e.g. tumours, white matter lesions), which can in turn provide biomarkers (e.g. the volume of the structure/lesion or other more sophisticated derived measures). Image registration is presented in Chapter 14. It is also a fundamental image analysis task which allows aligning images from different modalities or different patients and which is a prerequisite for many other ML methods. Chapter 15 describes methods for computer-aided diagnosis and prediction. These include methods to automatically classify patients (for instance to assist diagnosis) as well as to predict their future state. Chapter 16 presents ML methods to discover disease subtypes. Indeed, brain disorders are heterogeneous and patients with a given diagnosis may have different symptoms, a different underlying pathophysiology and a different evolution. Such heterogeneity is a major barrier to the development of new treatments. ML has the potential to help discover more homogeneous disease subtypes. Modelling disease progression is the focus of Chapter 17. The chapter describes a wide range of techniques that allow, in a data-driven manner, to build models of disease progression, which includes finding the ordering by which different biomarkers become abnormal, estimating trajectories of change and uncovering different evolution profiles within a given population. Chapter 18 is devoted to computational pathology which is the automated analysis of histological data (which may come from biopsies or post-mortem samples). Tremendous progresses have been made in this area in the past years. Chapter 19 describes methods for integrating multimodal data including medical imaging, clinical data and genetics (or other omics data). Indeed, characterizing the complexity of brain disorders requires to integrate multiple types of data, but such integration raises computational challenges.

Part IV is dedicated to validation and datasets. These are fundamental issues that are sometimes overlooked by ML researchers. It is indeed crucial that ML models for medicine are thoroughly and rigorously validated. Chapter 20 covers model validation. It introduces the main performance metrics for classification and regression tasks, describes how to estimate these metrics in an unbiased manner and how to obtain confidence intervals. Chapter 21 deals with reproducibility, the ability to reproduce results and findings. It is widely recognized that many fields of science, including ML for medicine, are undergoing a reproducibility crisis. The chapter describes the main types of reproducibility, what they require and why they are important. The topic of Chapter 22 is interpretability of ML methods. In particular, it reviews the main approaches to get insight on how “black-box” models take their decisions and describes their application to brain imaging data. Chapter 23 provides a regulatory science perspective on performance assessment of ML algorithms. It is indeed crucial to understand such perspective because regulation is critical to translate safe and effective technologies to the clinic. Finally, Chapter 24 provides an overview of the main existing datasets accessible to researchers. It can help scientists identify which datasets are most suited to a particular research question and provides hints on how to use them.

Part V presents applications of ML to various neurological and psychiatric disorders. Each chapter is devoted to a specific disorder or family of disorders. It presents some information about the disorder that should, in particular, be useful to researchers who don’t have a medical background. It then describes some important applications of ML to

this disorder as well as future challenges. The following disorders are covered: Alzheimer's disease and related dementia (including vascular dementia, frontotemporal dementia and dementia with Lewy bodies) in Chapter 25, Parkinson's disease and related disorders (including multiple system atrophy, progressive supranuclear palsy and dementia with Lewy bodies) in Chapter 26, epilepsy in Chapter 27, multiple sclerosis in Chapter 28, cerebrovascular disorders (including stroke, microbleeds, vascular malformations, aneurysms and small vessel disease) in Chapter 29, brain tumours in Chapter 30, neurodevelopmental disorders (including autism spectrum and attention deficit with hyperactivity disorders) in Chapter 31 and psychiatric disorders (including depression, schizophrenia and bipolar disorder) in Chapter 32.

We hope that this book will serve as a reference for researchers and graduate students who are new to this field of research as well as constitute a useful resource for all scientists working in this exciting scientific area.

Paris, France

Olivier Colliot

Acknowledgements

I would like to express my profound gratitude to all authors for their contributions to the book. It is thanks to you that this book has become a reality. I am also extremely grateful to all present and past members of the ARAMIS team. Research is a collective endeavour. It was a privilege (and a pleasure!) to work with you throughout all these years. I learn everyday thanks to you all. More generally, I would like to acknowledge all colleagues with whom I had the chance to work. I warmly thank the reviewers who have kindly reviewed chapters: Maria Gloria Bueno García, Sarah Cohen-Boulakia, Renaud David, Guillaume Dumas, Anton Iftimovici, Hicham Janati, Jochen Klucken, Margy McCullough-Hicks, Paolo Missier, Till Nicke, Sebastian Raschka, Denis Schwartz as well as those who preferred to stay anonymous. Your comments were very useful in improving the book. Finally, I would like to thank my family members for their love and support.

I acknowledge the following funding sources: the French government under management of the Centre National de la Recherche Scientifique, the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6), the European Union H2020 program (project EuroPOND, grant number 666992), the Paris Brain Institute under the Big Brain Theory Program (project PredictICD, project IMAGIN-DEAL in MS), Inria under the Inria Project Lab Program (project Neuromarkers), the Abeona Foundation (project Brain@Scale) and the Fondation Vaincre Alzheimer (grant number FR-18006CB). This book was made Open Access thanks to the support of the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

Paris, France

Olivier Colliot

Abbreviations

ABC	Activities-Specific Balance Confidence
ABCD	Adolescent Brain Cognitive Development
ACR	American College of Radiology
AD	Alzheimer's Disease
Adagrad	Adaptive Gradient optimizer
ADAS	Alzheimer's Disease Assessment Scale
ADC	Apparent Diffusion Coefficient
ADHD	Attention Deficit Hyperactivity Disorder
ADNI	Alzheimer's Disease Neuroimaging Initiative
AE	Autoencoder
AI	Artificial Intelligence
AIBL	Australian Imaging Biomarkers and Lifestyle Study of Aging
ALS	Amyotrophic Lateral Sclerosis
AP-HP	Assistance Publique-Hôpitaux de Paris
ARDM	Auto-Regressive Diffusion Models
ASD	Autism Spectrum Disorder
ASL	Arterial Spin Labelling
ASNR	American Society of Neuroradiology
ASPECTS	The Alberta Stroke Programme Early CT Score
ASR	Age-Standardized incidence Rate
ASSD	Average Symmetric Surface Distance
ATP	Adenosine Triphosphate
AUC	Area Under the Curve (can apply to ROC curve or PR curve for instance)
AVM	Arteriovenous Malformation
BA	Balanced Accuracy
BAM	Binary Alignment Map
BATS	Brisbane Adolescent Twin Study
BCI	Brain Computer Interface
BD	Bipolar Disorder
Bi-LSTM	Bi-directional Long Short-Term Memory
BMI	Body Mass Index
BOLD	Blood-Oxygen-Level-Dependent
BPTT	Back Propagation Through Time
BraTS	Brain Tumour Segmentation Challenge
BRNN	Bi-directional Recurrent Neural Network
CA	Cornu Ammonis
CA	Cross-Attention
CAD	Computer-Assisted Diagnosis
CADASIL	Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy
CAM	Class Activation Maps
CAP	College of American Pathologists
CARS	Coherent Anti-Stokes Raman Scattering
CBD	Cortico-basal Degeneration

CBF	Cerebral Blood Flow
CBV	Cerebral Blood Volume
CC	Cross-Correlation
CCA	Canonical Correlation Analysis
CDF	Cumulative Density Function
CDR	Clinical Dementia Rating
CDRH	Center for Devices and Radiological Health at FDA
cGAN	Conditional Generative Adversarial Network
CIFAR	Canadian Institute For Advanced Research
CIS	Clinically Isolated Syndrome
CLAIM	Checklist for Artificial Intelligence in Medical imaging
CLIP	Contrastive Language-Image Pretraining
C-LSTM	Convolutional Long Short-Term Memory
CMB	Cerebral Microbleed
CN	Healthy Controls (or Cognitively Normal participants)
CNN	Convolutional Neural Network
CNS	Central Nervous System
CNV	Copy number variant
CP	Computational Pathology
CPAB	Continuous Piecewise Affine-Based
CPM	Computational Precision Medicine
CPM-RadPath	CPM Radiology-Pathology Challenge
CPRD	Clinical Practice Research Datalink
CRAM	Compressed Reference-oriented Alignment Map
CSF	Cerebrospinal Fluid
cSVD	Cerebral Small Vessel Disease
CT	Computed Tomography
CTA	Computed Tomography Angiography
CTSA	Clinical Translational Science Awards
CTV-3	Clinical Terms Version 3
CV	Cross-Validation
D3PM	Data-Driven Disease Progression Modelling
DAT	Dopamine Transporter
dbGAP	Database of Genotypes and Phenotypes
DBS	Deep Brain Stimulation
DCGAN	Deep Convolutional Generative Adversarial Network
DDPM	Denoising Diffusion Probabilistic Models
DEBM	Discriminative Event-Based Model
DIAN	Dominantly Inherited Alzheimer Network
DICOM	Digital Imaging and Communications in medicine
DIR	Double Inversion Recovery MR sequence
DL	Deep Learning
DLB	Dementia with Lewy Bodies
DNA	Deoxyribonucleic Acid
dof	Degrees of Freedom
DPCE	Distance map Penalized Cross Entropy loss
DPS	Disease Progression Score
DRAM	Deep Recurrent Attention Model
DRNN	Disconnected Recurrent Neural Network
DSA	Digital Subtraction Angiography
DSC	Dice Similarity Coefficient

DSM-5	5th edition of the Diagnostic and Statistical Manual of Mental Disorders
DT	Decision Tree
DTI	Diffusion Tensor Imaging
DWI	Diffusion-Weighted Imaging
DXA	Dual-energy X-ray Absorptiometry
DZ	Dizygotic
EBM	Event-Based Model
EDSS	Expanded Disability Status Scale
EEA	European Economic Area
EEG	Electroencephalography
EFA	Exploratory Factor Analysis
EHR	Electronic Health Record
ELBO	Evidence Lower Bound
ELL	Exponential Logarithmic Loss
EM	Expectation-Maximization
ENIGMA	Enhancing NeuroImaging Genetics Through Meta-Analysis
eQTL	expression Quantitative Trait Loci
ET	Enhancing Tumour
EU	European Union
FA	Fractional Anisotropy
FAIR	Findable, Accessible, Interoperable, Reusable
FASTA	Fast-All format
FCD	Focal Cortical Dysplasia
FCN	Fully Connected Network
FCNN	Fully Convolutional Neural Network
FDA	United States Food and Drug Administration
FDG-PET	[¹⁸ F]-Fluorodeoxyglucose Positron Emission Tomography
FDR	False Discovery Rate
FFPE	Formalin-Fixed Paraffin-Embedded
FID	Fréchet Inception Distance
FLAIR	Fluid-Attenuated Inversion Recovery
FLOP	Floating Point Operations
fMRI	functional Magnetic Resonance Imaging
FN	False Negative
FOG	Freezing of Gait
FP	False Positive
FROC	Free-response ROC
FTLD	Fronto-temporal Lobar Degeneration
G2PSR	Genome-to-Phenome Sparse Regression
GAN	Generative Adversarial Network
GBM	Glioblastoma
GD	Generalized Dice loss
GDPR	General Data Protection Regulation
GENFI	Genetic FTD Initiative
GEO	Gene Expression Omnibus
GFF	General Feature Format
GIS	Geographic Information Systems
GM	Gray Matter
GMM	Gaussian Mixture Model

GO	Gene Ontology
GPPM	Gaussian Process Progression Model
GPS	Global Positioning System
GPU	Graphical Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
GRE	Gradient-Recalled Echo
GRU	Gated Recurrent Unit
GTEx	Genotype-Tissue Expression
GWAS	Genome-Wide Association Study
H&E	Hematoxylin and Eosin
HAR	Human Activity Recognition
HC	Healthy Controls
HCP-YA	Human Connectome Project Young Adult
HD	Hausdorff Distance
HGG	High-Grade Glioma
HIPAA	Health Insurance Portability and Accountability Act
HS	Hippocampal Sclerosis
HUPO	Human Proteome Organization
HYDRA	Heterogeneity through Discriminative Analysis
i.i.d.	Independent and Identically Distributed
IA	Intracranial Aneurysm
ICA	Independent Component Analysis
ICD	International Classification of Diseases
iCDF	Inverse Cumulative Density Function
ICF	International Classification of Functioning, Disability and Health
ID	Intelligence Disabilities
IDH	Isocitrate Dehydrogenase
IEEE	Institute of Electrical and Electronics Engineers
IHC	Inmunohistochemistry
IoU	Intersection over Union also called JI
IPMI	Information Processing in Medical Imaging conference
IQ	Intelligence Quotient
iRANO	Immune-related Response Assessment in Neuro-Oncology
iRBD	Idiopathic Rapid eye movement sleep Behaviour Disorder
ISBI	International Symposium on Biomedical Imaging
ISLES	The Ischemic Stroke Lesion Segmentation
JI	Jaccard index also called IoU
JS/JSD	Jensen-Shannon Divergence
KDE	Kernel Density Estimate
KDE-EBM	Kernel Density Estimation EBM
KEGG	Kyoto Encyclopedia of Genes and Genomes
KL/KLD	Kullback-Leibler Divergence
kNN	<i>k</i> -Nearest Neighbours
LASSO	Least Absolute Shrinkage and Selection Operator
LATE	Limbic-predominant Age-related TDP-43 Encephalopathy
LDA	Linear Discriminant Analysis
LDDMM	Large Deformation Diffeomorphic Metric Mapping
LGG	Low-Grade Glioma
LIME	Local Interpretable Model-agnostic Explanations

LJTMM	Latent Time Joint Mixed Model
LP	Linear Programming
LPA	Logopenic Progressive Aphasia
LR	Logistic Regression
LR-	Negative Likelihood Ratio
LR+	Positive Likelihood Ratio
LRP	Layer-wise relevance
LSTM	Long Short-Term Memory
LVO	Large Vessel Occlusion
MAE	Mean Absolute Error
MAGIC	Multi-scAle heterogeneity analysIs and Clustering
MAGNIMS	Magnetic Resonance Imaging in Multiple Sclerosis network
MAP	Maximum a Posteriori
MAR	Missing at Random
MCA	Multi-head Cross-Attention
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
mcVAE	Multi-Channel Variational Autoencoder
MD	Mean Diffusivity
MDD	Major Depressive Disorder
MDE	Major Depressive Episode
MDS-UPDRS	Movement Disorder Society Unified Parkinson's Disease Rating Scale (synonymous: UPDRS)
MEDA	Minimal Evidence of Disease Activity
MEG	Magnetoencephalography
MEM	Micro Electro Mechanical system
MGMT	O6-Methylguanine-DNA Methyltransferase
MI	Mutual Information
MICCAI	The Medical Image Computing and Computer Assisted Intervention Society
MICE	Multiple Imputation by Chained Equations
MIDS	Medical Imaging Data Structure
MIP	Maximal Intensity Projection
ML	Machine Learning
MLE	Maximum Likelihood Estimation
MLP	Multi-Layer Perceptron
MMSA	Masked Multi-head Self-Attention
MMSE	Mini-Mental state examination
MNAR	Missing Not at Random
MND	Motor Neuron Disease
MNI	Montreal Neurological Institute
MNIST	Modified National Institute of Standards and Technology dataset
MoCA	Montreal Cognitive Assessment
MRA	Magnetic Resonance Angiography
MRI	Magnetic Resonance Imaging
mRNA	Messenger RNA
mRS	modified Rankin Score
MS	Multiple Sclerosis
MSA	Multi-head Self-Attention

MSA	Multiple System Atrophy
MSA-C	Cerebellar variant of Multiple System Atrophy
MSA-P	Parkinsonian variant of Multiple System Atrophy
MSD	Medical Segmentation Decathlon
MSE	Mean Squared Error
mTOR	Mammalian Target of Rapamycin
MTR	Magnetization Transfer Ratio
MZ	Monozygotic
mzML	Mass Spectrometry Markup Language
NAWM	Normal Appearing White Matter
NB	Naive Bayes
NCBI	National Center for Biotechnology Information
NCC	Normalized Cross Correlation
ncRNA	Non-coding RNA
NDDs	Neurodevelopmental Disorders
NEDA	No Evidence of Disease Activity
NeurIPS	Neuronal Information Processing Systems conference
NGS	Next Generation Sequencing
NifTI	Neuroimaging Informatics Technology Initiative
NIH	National Institutes of Health
NINCDS-ADRDA	National Institute of Neurological and Communicative Disorders and Stroke - Alzheimer's Disease and Related Disorders Association
NIPALS	Non-linear Iterative Partial Least Squares
NIVEL	Netherlands Institute for Health Services Research
NIVEL-PCD	NIVEL Primary Care Database
NLP	Natural Language Processing
NMF	Non-negative Matrix Factorization
NMI	Normalized Mutual Information
NMOSD	Neuromyelitis Optica Spectrum Disorder
NMT	Neural Machine Translation
NN	Neural Network
NPV	Negative Predictive Value
OATS	Older Adult Twin Study
OCD	Obsessive Compulsive Disorder
OCT	Optimal Cutting Temperature
OSF	Open Science Framework
PACS	Picture Archiving and Communication System
PCA	Posterior Cortical Atrophy
PCA	Principal Component Analysis
PD	Parkinson's Disease
PD	Proton-Density MR sequence
PDF	Probability Density Function
PE	Positional Encoding
PET	Positron Emission Tomography
PIB-PET	[¹¹ C]-Pittsburgh Compound B Positron Emission Tomography
PiD	Pick's Disease
PLS	Partial Least Squares
PLSR	Partial Least Square Regression
PML	Progressive Multifocal Leukoencephalopathy

PNS	Peripheral Nervous System
PPA	Primary Progressive Aphasia
PPMI	Parkinson's Progression Markers Initiative
PPMS	Primary Progressive Multiple Sclerosis
PPV	Positive Predictive Value
PRS	Polygenic Risk Score
PSI	HUPO Proteomics Standards Initiative
PSI	Proteomics Standards Initiative
PSP	Progressive Supranuclear Palsy
psPD	pseudoprogression of disease
PT	Patient
PTSD	Post-Traumatic Stress Disorder
PVS	Perivascular Space
PWI	Perfusion Weighted Imaging
QSM	Quantitative Susceptibility Mapping
QT	Quantitative Traits
QTAB	Queensland Twin Adolescent Brain
QTIM	Queensland Twin IMaging
rΔCBF	relative CBF change
RANO	Response Assessment in Neuro-Oncology
RAVEL	Removal of Artificial Voxel Effect by Linear regression
RBF	Radial Basis Function
RCNN	Region Convolutional Neural Network
RECIST	Response Evaluation Criteria in Solid Tumours
ReLU	Rectified Linear Unit
REM	Rapid Eye Movement
ResNet	Residual Neural Network
RF	Random Forest
RKHS	Reproducing Kernel Hilbert Space
RMSE	Root Mean Square Error
RMSProp	Root Mean Squared Propagation
RNA	Ribonucleic Acid
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic curve
RRMS	Relapsing Remitting Multiple Sclerosis
RS-fMRI	Resting State functional Magnetic Resonance Imaging
RSNA	Radiological Society of North America
SA	Self-Attention
SAM	Sequence Alignment Map
SD	Standard Deviation
SDG	Stochastic Gradient Descent
SHAP	SHapley Additive exPlanations
Smile-GAN	Semi-supervised cLustering via GANs
SNOMED-CT	Systematized NOmenclature of MEDicine - Clinical Terms
SNP	Single Nucleotide Polymorphism
SPECT	Single-Photon Emission Computed Tomography
SPIE	The Society for Photoelectrical Instrumentation Engineers - The International Society for Optics and Photonics
SPIRIT-AI	Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence

SPMS	Secondary Progressive Multiple Sclerosis
SRA	Sequence Read Archive
SRH	Stimulated Raman scattering Histology
SS	Sensitivity-Specificity loss
SSD	Sum of Square Differences
SSL	Semi-Supervised Learning
STARD-AI	Standards for Reporting Diagnostic Accuracy Studies - Artificial Intelligence
STN	Spatial Transformer Network
STR	Swedish Twin Registry
STRIVE	the STandards for ReportIng Vascular changes on nEuroimaging
SUD	Substance Use Disorder
SuLign	Subtyping Alignment
SuStaIn	Subtype and Stage Inference
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SWI	Susceptibility-Weighted Images
TC	Tumour Core
TEBM	Temporal Event-Based Model
TICI	Thrombolysis in Cerebral Infarction
TLE	Temporal Lobe Epilepsy
TMZ	Temozolomide
TN	True Negative
TNR	True Negative Rate
TOPMed	Trans-omics Precision Medicine
TP	True Positive
TPR	True Positive Rate
TRIPOD-ML	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-Machine Learning
tRNA	Transfer RNA
UAD	Unsupervised Anomaly Detection
UDA	Unsupervised Data Augmentation
UI	User Interface
UKB	UK Biobank
UMLS	Unified Medical Language System
UPDRS	Unified Parkinson's Disease Rating Scale
UX	User Experience
VaD	Vascular Dementia
VAE	Variational Autoencoder
VA-GAN	Visual Attribution Generative Adversarial Network
VASARI	Visually AcceSAble Rembrandt Images
VCCA	Deep Variational CCA
VCI	Vascular Cognitive Impairment
VETSA	Vietnam Era Twin Study of Aging
ViT	Vision Transformer
ViViT	Video Vision Transformer
VQGAN	Vector Quantization Generative Adversarial Network
VQ-VAE	Vector Quantization Variational Autoencoder
WCE	Weighted Cross Entropy loss

WGAN	Wasserstein Generative Adversarial Network
WGS	Whole Genome Sequence
WHO	World Health Organization
WM	White Matter
WMH	White Matter Hyperintensity
WSI	Whole Slide Image
WT	Whole Tumour
WUSTL	Washington University in Saint Louis
xAI	eXplainable AI
XML	eXtensible Markup Language
XNAT	eXtensible Neuroimaging Archive Toolkit

Contents

<i>Preface to the Series</i>	v
<i>Preface</i>	vii
<i>Acknowledgements</i>	xii
<i>Abbreviations</i>	xiii
<i>Contributors</i>	xxvii

PART I MACHINE LEARNING FUNDAMENTALS

1 A Non-technical Introduction to Machine Learning..... <i>Olivier Colliot</i>	3
2 Classic Machine Learning Methods <i>Johann Faouzi and Olivier Colliot</i>	25
3 Deep Learning: Basics and Convolutional Neural Networks (CNNs)..... <i>Maria Vakalopoulou, Stergios Christodoulidis, Ninon Burgos, Olivier Colliot, and Vincent Lepetit</i>	77
4 Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Introduction to Influential Research..... <i>Susmita Das, Amara Tariq, Thiago Santos, Sai Sandeep Kantareddy, and Imon Banerjee</i>	117
5 Generative Adversarial Networks and Other Generative Models	139
<i>Markus Wenzel</i>	
6 Transformers and Visual Transformers..... <i>Robin Courant, Maika Edberg, Nicolas Dufour, and Vicky Kalogeiton</i>	193

PART II DATA

7 Clinical Assessment of Brain Disorders..... <i>Stéphane Epelbaum and Federica Cacciamani</i>	233
8 Neuroimaging in Machine Learning for Brain Disorders..... <i>Ninon Burgos</i>	253
9 Electroencephalography and Magnetoencephalography..... <i>Marie-Constance Corsi</i>	285
10 Working with Omics Data: An Interdisciplinary Challenge at the Crossroads of Biology and Computer Science..... <i>Thibault Poinsignon, Pierre Poulain, Mélina Gallopin, and Gaëlle Lelandais</i>	313
11 Electronic Health Records as Source of Research Data	331
<i>Wenjuan Wang, Davide Ferrari, Gabriel Haddon-Hill, and Vasa Curcin</i>	

- 12 Mobile Devices, Connected Objects, and Sensors 355
*Sirenia Lizbeth Mondragón-González, Eric Burguière,
and Karim N'diaye*

PART III METHODOLOGIES

- 13 Medical Image Segmentation Using Deep Learning 391
Han Liu, Dewei Hu, Hao Li, and Ipek Oguz
- 14 Image Registration: Fundamentals and Recent Advances
Based on Deep Learning 435
Min Chen, Nicholas J. Tustison, Robit Jena, and James C. Gee
- 15 Computer-Aided Diagnosis and Prediction in Brain Disorders 459
*Vikram Venkatraghavan, Sebastian R. van der Voort,
Daniel Bos, Marion Smits, Frederik Barkhof, Wiro J. Niessen,
Stefan Klein, and Esther E. Bron*
- 16 Subtyping Brain Diseases from Imaging Data 491
*Junhao Wen, Erdem Varol, Zhijian Yang, Gyujoon Hwang,
Dominique Dwyer, Anahita Fathi Kazerooni,
Paris Alexandros Lalousis, and Christos Davatzikos*
- 17 Data-Driven Disease Progression Modeling 511
Neil P. Oxtoby
- 18 Computational Pathology for Brain Disorders 533
Gabriel Jiménez and Daniel Racoceanu
- 19 Integration of Multimodal Data 573
*Marco Lorenzi, Marie Deprez, Irene Balelli,
Ana L. Aguila, and Andre Altmann*

PART IV VALIDATION AND DATASETS

- 20 Evaluating Machine Learning Models and Their Diagnostic Value 601
Gael Varoquaux and Olivier Colliot
- 21 Reproducibility in Machine Learning for Medical Imaging 631
Olivier Colliot, Elina Thibeau-Sutre, and Ninon Burgos
- 22 Interpretability of Machine Learning Methods Applied
to Neuroimaging 655
*Elina Thibeau-Sutre, Sasha Collin, Ninon Burgos,
and Olivier Colliot*
- 23 A Regulatory Science Perspective on Performance Assessment
of Machine Learning Algorithms in Imaging 705
*Weijie Chen, Daniel Krainak, Berkman Sahiner,
and Nicholas Petrick*

- 24 Main Existing Datasets for Open Brain Research on Humans 753
*Baptiste Couvy-Duchesne, Simona Bottani, Etienne Camenen,
Fang Fang, Mulusew Fikere, Juliana Gonzalez-Astudillo, Joshua Harvey,
Ravi Hassanaly, Irfahan Kassam, Penelope A. Lind, Qianwei Liu, Yi Lu,
Marta Nabais, Thibault Rolland, Julia Sidorenko, Lachlan Strike, and
Margie Wright*

PART V DISORDERS

- 25 Machine Learning for Alzheimer’s Disease and Related Dementia 807
*Marc Modat, David M. Cash, Liane Dos Santos Canas,
Martina Bocchetta, and Sébastien Ourselin*
- 26 Machine Learning for Parkinson’s Disease and Related Disorders 847
Johann Faouzi, Olivier Colliot, and Jean-Christophe Corvol
- 27 Machine Learning in Neuroimaging of Epilepsy 879
*Hyo Min Lee, Ravnoor Singh Gill, Neda Bernasconi,
and Andrea Bernasconi*
- 28 Machine Learning in Multiple Sclerosis 899
Bas Jaspertse and Frederik Barkhof
- 29 Machine Learning for Cerebrovascular Disorders 921
Yannan Yu and David Yen-Ting Chen
- 30 The Role of Artificial Intelligence in Neuro-oncology Imaging 963
*Jennifer Soum, Lu-Aung Yosuke Masudathaya,
Arabdha Biswas, and Daniel S. Chow*
- 31 Machine Learning for Neurodevelopmental Disorders 977
Clara Moreau, Christine Deruelle, and Guillaume Auzias
- 32 Machine Learning and Brain Imaging for Psychiatric Disorders:
New Perspectives 1009
Ivan Brossolle, Quentin Gallet, Pauline Favre, and Josselin Houenou
- Disclosure Statement of the Editor* 1037
Index 1039

Contributors

- ANA L. AGUILA • *University College London, Centre for Medical Image Computing, COMBINE Lab, London, UK*
- ANDRE ALTMANN • *University College London, Centre for Medical Image Computing, COMBINE Lab, London, UK*
- GUILLAUME AUZIAS • *Aix-Marseille Université, CNRS, Institut de Neurosciences de la Timone, Marseille, France*
- IRENE BALELLI • *Université Côte d'Azur, Inria Sophia Antipolis, Epione Research Group, Nice, France*
- IMON BANERJEE • *Mayo Clinic, Phoenix, AZ, USA; Arizona State University, School of Computing, Informatics, and Decision Systems Engineering, Tempe, AZ, USA*
- FREDERIK BARKHOF • *Department of Radiology and Nuclear Medicine, Amsterdam University Medical Center, Amsterdam, The Netherlands; Queen Square Institute of Neurology and Centre for Medical Image Computing, University College, London, UK*
- ANDREA BERNASCONI • *McGill University, Montreal Neurological Institute and Hospital, Montreal, QC, Canada*
- NEDA BERNASCONI • *McGill University, Montreal Neurological Institute and Hospital, Montreal, QC, Canada*
- ARABDHA BISWAS • *Department of Radiological Sciences, University of California, Irvine, Irvine, CA, USA*
- MARTINA BOCCHETTA • *UCL Queen Square Institute of Neurology, Dementia Research Centre, London, UK; Centre for Cognitive and Clinical Neuroscience, Division of Psychology, Department of Life Sciences, College of Health, Medicine and Life Sciences, Brunel University, London, UK*
- DANIEL BOS • *Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands; Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands*
- SIMONA BOTTANI • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France*
- ESTHER E. BRON • *Biomedical Imaging Group Rotterdam, Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands*
- IVAN BROSSOLLET • *Neurospin, UNIACIT Lab, PsyBrain Team, CEA Saclay, Gif-sur-Yvette, France*
- NINON BURGOS • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France*
- ERIC BURGUIÈRE • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France*
- FEDERICA CACCIAMANI • *University of Bordeaux, Inserm, UMR Bordeaux Population Health, PHARes Team, Bordeaux, France*
- ETIENNE CAMENEN • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France*
- DAVID M. CASH • *UCL Queen Square Institute of Neurology, Dementia Research Centre, London, UK; UK Dementia Research Institute at UCL, London, UK*
- DAVID YEN-TING CHEN • *Department of Medical Imaging, Taipei Medical University—Shuang Ho Hospital, Zhonghe District, New Taipei City, Taiwan*

- MIN CHEN • *University Pennsylvania, Department of Radiology, Philadelphia, PA, USA*
- WEIJIE CHEN • *Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD, USA*
- DANIEL S. CHOW • *Department of Radiological Sciences, University of California, Irvine, Irvine, CA, USA*
- STERGIOS CHRISTODOULIDIS • *Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes, Gif-sur-Yvette, France*
- SASHA COLLIN • *Sorbonne Université, Institut du Cerveau – Paris Brain Institute – ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, Paris, France*
- OLIVIER COLLIOT • *Sorbonne Université, Institut du Cerveau – Paris Brain Institute – ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, Paris, France*
- MARIE-CONSTANCE CORSI • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France*
- JEAN-CHRISTOPHE CORVOL • *Sorbonne Université, Institut du Cerveau – Paris Brain Institute – ICM, CNRS, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, Department of Neurology, Paris, France*
- ROBIN COURANT • *LIX, CNRS, Ecole Polytechnique, IP Paris, Paris, France; CNRS, IRISA, INRIA, Univ. Rennes, Rennes, France*
- BAPTISTE COUVY-DUCHESNE • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France; Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD, Australia*
- VASA CURCIN • *Department of Population Health Sciences, King's College London, London, United Kingdom*
- SUSMITA DAS • *Indian Institute of Technology (IIT), Centre of Excellence in Artificial Intelligence, Kharagpur, West Bengal, India*
- CHRISTOS DAVATZIKOS • *Center for Biomedical Image Computing and Analytics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
- MARIE DEPREZ • *Université Côte d'Azur, Inria Sophia Antipolis, Epione Research Group, Nice, France*
- CHRISTINE DERUELLE • *Aix-Marseille Université, CNRS, Institut de Neurosciences de la Timone, Marseille, France*
- LIANE DOS SANTOS CANAS • *King's College London, School of Biomedical Engineering & Imaging Sciences, London, UK*
- NICOLAS DUFOUR • *LIX, CNRS, Ecole Polytechnique, IP Paris, Paris, France*
- DOMINIQUE DWYER • *Department of Psychiatry and Psychotherapy, Ludwig-Maximilian University, Munich, Germany*
- MAIKA EDBERG • *LIX, CNRS, Ecole Polytechnique, IP Paris, Paris, France*
- STÉPHANE EPELBAUM • *Sorbonne Université, Institut du Cerveau – Paris Brain Institute – ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, Paris, France; AP-HP, Hôpital de la Pitié-Salpêtrière, Department of Neurology, Institut de la Mémoire et de la Maladie d'Alzheimer (IM2A), Paris, France*
- FANG FANG • *Karolinska Institutet (KI), Stockholm, Sweden*
- JOHANN FAOUZI • *CREST, ENSAI, Campus de Ker-Lann, Bruz Cedex, France; Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France*

- PAULINE FAVRE • *Neurospin, UNIACIT Lab, PsyBrain Team, CEA Saclay, Gif-sur-Yvette, France; INSERM U955, Translational Neuropsychiatry Team, Faculté de Santé, Université Paris Est Créteil, Créteil, France*
- DAVIDE FERRARI • *Department of Population Health Sciences, King's College London, London, United Kingdom*
- MULUSEW FIKERE • *Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD, Australia*
- QUENTIN GALLET • *Neurospin, UNIACIT Lab, PsyBrain Team, CEA Saclay, Gif-sur-Yvette, France; INSERM U955, Translational Neuropsychiatry Team, Faculté de Santé, Université Paris Est Créteil, Créteil, France*
- MÉLINA GALLOPIN • *Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Saclay, Gif-sur-Yvette, France*
- JAMES C. GEE • *University Pennsylvania, Department of Radiology, Philadelphia, PA, USA*
- RAVNOOR SINGH GILL • *McGill University, Montreal Neurological Institute and Hospital, Montreal, QC, Canada*
- JULIANA GONZALEZ-ASTUDILLO • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France*
- GABRIEL HADDON-HILL • *Department of Population Health Sciences, King's College London, London, United Kingdom*
- JOSHUA HARVEY • *University of Exeter Medical School, RILD Building, RD&E Hospital Wonford, Exeter, UK*
- RAVI HASSANALY • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France*
- JOSSELIN HOUENOU • *Neurospin, UNIACIT Lab, PsyBrain Team, CEA Saclay, Gif-sur-Yvette, France; INSERM U955, Translational Neuropsychiatry Team, Faculté de Santé, Université Paris Est Créteil, Créteil, France; APHP, Mondor Univ. Hospitals, DMU Impact, Psychiatry Department, Créteil, France*
- DEWEI HU • *Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, USA*
- GYUJOON HWANG • *Center for Biomedical Image Computing and Analytics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
- BAS JASPERSE • *Department of Radiology and Nuclear Medicine, Amsterdam University Medical Center, Amsterdam, The Netherlands*
- ROHIT JENA • *University Pennsylvania, Department of Radiology, Philadelphia, PA, USA*
- GABRIEL JIMÉNEZ • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France*
- VICKY KALOGEITON • *LIX, CNRS, Ecole Polytechnique, IP Paris, Paris, France*
- SAI SANDEEP KANTAREDDY • *Arizona State University, School of Computing, Informatics, and Decision Systems Engineering, Tempe, AZ, USA*
- IRFAHAN KASSAM • *Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore*
- ANAHITA FATHI KAZEROONI • *Institute for Mental Health and Centre for Human Brain Health, School of Psychology, University of Birmingham, Birmingham, UK*
- STEFAN KLEIN • *Biomedical Imaging Group Rotterdam, Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands*

- DANIEL KRAINAK • *Division of Radiological Health, Office of In Vitro Diagnostics and Radiological Health, Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD, USA*
- PARIS ALEXANDROS LALOUSIS • *Institute for Mental Health and Centre for Human Brain Health, School of Psychology, University of Birmingham, Birmingham, UK*
- HYO MIN LEE • *McGill University, Montreal Neurological Institute and Hospital, Montreal, QC, Canada*
- GAËLLE LELANDAIS • *Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Saclay, Gif-sur-Yvette, France*
- VINCENT LEPESTIT • *LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France*
- HAO LI • *Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, USA*
- PENELOPE A. LIND • *Psychiatric Genetics, QIMR Berghofer Medical Research Institute, Herston, QLD, Australia; School of Biomedical Sciences, Queensland University of Technology, Kelvin Grove, QLD, Australia; School of Biomedical Sciences, Faculty of Medicine, University of Queensland, St Lucia, QLD, Australia*
- HAN LIU • *Department of Computer Science, Vanderbilt University, Nashville, TN, USA*
- QIANWEI LIU • *Karolinska Institutet (KI), Stockholm, Sweden*
- MARCO LORENZI • *Université Côte d'Azur, Inria Sophia Antipolis, Epione Research Group, Nice, France*
- YI LU • *Karolinska Institutet (KI), Stockholm, Sweden*
- LU-AUNG YOSUKE MASUDATHAYA • *Department of Radiological Sciences, University of California, Irvine, Irvine, CA, USA*
- MARC MODAT • *King's College London, School of Biomedical Engineering & Imaging Sciences, London, UK*
- SIRENIA LIZBETH MONDRAGÓN-GONZÁLEZ • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, Paris, France*
- CLARA MOREAU • *Human Genetics and Cognitive Functions, CNRS UMR 3571, Université de Paris, Institut Pasteur, Paris, France*
- KARIM N'DIAYE • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, Paris, France*
- MARTA NABAIS • *University of Exeter Medical School, RILD Building, RD&E Hospital Wonford, Exeter, UK*
- WIRO J. NIJSEN • *Biomedical Imaging Group Rotterdam, Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands; Quantitative Imaging Group, Department of Imaging Physics, Faculty of Applied Sciences, TU Delft, The Netherlands*
- IPEK OGUZ • *Department of Computer Science, Vanderbilt University, Nashville, TN, USA; Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, USA*
- SÉBASTIEN OURSELIN • *King's College London, School of Biomedical Engineering & Imaging Sciences, London, UK*
- NEIL P. OXTOBY • *UCL Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK*
- NICHOLAS PETRICK • *Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD, USA*

- THIBAULT POINSIGNON • *Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Saclay, Gif-sur-Yvette, France*
- PIERRE POULAIN • *Université Paris Cité, CNRS, Institut Jacques Monod, Paris, France*
- DANIEL RACOCEANU • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France*
- THIBAULT ROLLAND • *Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France*
- BERKMAN SAHINER • *Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD, USA*
- THIAGO SANTOS • *Emory University, Department of Computer Science, Atlanta, GA, USA*
- JULIA SIDORENKO • *Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD, Australia*
- MARION SMITS • *Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands*
- JENNIFER SOUN • *Department of Radiological Sciences, University of California, Irvine, Irvine, CA, USA*
- LACHLAN STRIKE • *Queensland Brain Institute, the University of Queensland, St Lucia, QLD, Australia*
- AMARA TARIQ • *Mayo Clinic, Phoenix, AZ, USA*
- ELINA THIBEAU-SUTRE • *Department of Applied Mathematics, Technical Medical Centre, University of Twente, Enschede, The Netherlands; Sorbonne Université, Institut du Cerveau – Paris Brain Institute – ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, Paris, France*
- NICHOLAS J. TUSTISON • *University of Virginia, Department of Radiology and Medical Imaging, Charlottesville, VA, USA*
- MARIA VAKALOPOULOU • *Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes, Gif-sur-Yvette, France*
- ERDEM VAROL • *Department of Statistics, Center for Theoretical Neuroscience, Zuckerman Institute, Columbia University, New York, NY, USA*
- GAEL VAROQUAUX • *Soda, Inria, Saclay, France*
- VIKRAM VENKATRAGHAVAN • *Alzheimer Center Amsterdam, Neurology, Vrije Universiteit, Amsterdam, The Netherlands; Amsterdam Neuroscience, Neurodegeneration, Amsterdam, The Netherlands*
- SEBASTIAN R. VAN DER VOORT • *Biomedical Imaging Group Rotterdam, Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands*
- WENJUAN WANG • *Department of Population Health Sciences, King's College London, London, United Kingdom*
- JUNHAO WEN • *Center for Biomedical Image Computing and Analytics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
- MARKUS WENZEL • *Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany*
- MARGIE WRIGHT • *Queensland Brain Institute, the University of Queensland, St Lucia, QLD, Australia; Centre for Advanced Imaging, The University of Queensland, St Lucia, QLD, Australia*
- ZHIJIAN YANG • *Center for Biomedical Image Computing and Analytics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
- YANNAN YU • *Department of Radiology, University of California San Francisco, San Francisco, CA, USA*

Part I

Machine Learning Fundamentals



Chapter 1

A Non-technical Introduction to Machine Learning

Olivier Colliot

Abstract

This chapter provides an introduction to machine learning for a non-technical readership. Machine learning is an approach to artificial intelligence. The chapter thus starts with a brief history of artificial intelligence in order to put machine learning into this broader scientific context. We then describe the main general concepts of machine learning. Readers with a background in computer science may skip this chapter.

Key words Machine learning, Artificial intelligence, Supervised learning, Unsupervised learning

1 Introduction

Machine learning (ML) is a scientific domain which aims at allowing computers to perform tasks without being explicitly programmed to do so [1]. To that purpose, the computer is trained using the examination of examples or experiences. It is part of a broader field of computer science called *artificial intelligence* (AI) which aims at creating computers with abilities that are characteristic of human or animal intelligence. This includes tasks such as perception (the ability to recognize images or sounds), reasoning, decision-making, or creativity. Emblematic tasks which are easy to perform for a human and are inherently difficult for a computer are, for instance, recognizing objects, faces, or animals in photographs or recognizing words in speech. On the other hand, there are also tasks which are inherently easy for a computer and difficult for a human, such as computing with large numbers or memorizing exactly huge amounts of text. Machine learning is the AI technique that has achieved the most impressive successes over the past years. However, it is not the only approach to AI, and conceptually different approaches also exist.

Machine learning also has close ties to other scientific fields. First, it has evident strong links to statistics. Indeed, most machine learning approaches exploit statistical properties of the data. Moreover, some classical approaches used in machine learning were

actually invented in statistics (for instance, linear or logistic regression). Nowadays, there is a constant interplay between progress in statistics and machine learning. ML has also important ties to signal and image processing, ML techniques being efficient for many applications in these domains and signal/image processing concepts being often key to the design or understanding of ML techniques. There are also various links to different branches of mathematics, including optimization and differential geometry. Besides, some inspiration for the design of ML approaches comes from the observation of biological cognitive systems, hence the connections with cognitive science and neuroscience. Finally, the term *data science* has become commonplace to refer to the use of statistical and computational methods for extracting meaningful patterns from data. In practice, machine learning and data science share many concepts, techniques, and tools. Nevertheless, data science puts more emphasis on the discovery of knowledge from the data, while machine learning focuses on solving tasks.

This chapter starts by providing a few historical landmarks regarding artificial intelligence and machine learning (Subheading 2). It then proceeds with the main concepts of ML which are foundational to understand other chapters of this book.

2 A Bit of History

As a scientific endeavor, artificial intelligence is at least 80 years old. Here, we provide a very brief overview of this history. For more details, the reader may refer to [2]. A non-exhaustive timeline of AI is shown in Fig. 1.

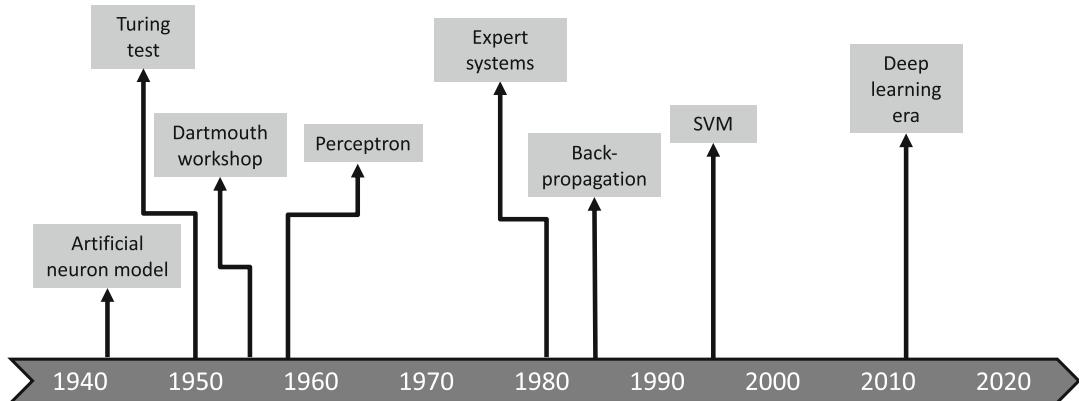


Fig. 1 A brief timeline of AI with some of the landmark advances

Even if this is debatable, one often considers AI to emerge in the 1940s–1950s with a series of important concepts and events. In 1943, the neurophysiologist Warren McCulloch and the logician Walter Pitts proposed an artificial neuron model, which is a mathematical abstraction of a biological neuron [3], and showed that sets of neurons can compute logical operations. In 1948, the mathematician and philosopher Norbert Wiener coined the term “cybernetics” [4] to designate the scientific study of control and communication in humans, animals, and machines. This idea that such processes can be studied within the same framework in both humans/animals and machines is a conceptual revolution. In 1949, the psychologist Donald Hebb [5] described a theory of learning for biological neurons which was later influential in the modification of the weights of artificial neurons.

In 1950, Alan Turing, one of the founders of computer science, introduced a test (the famous “Turing test”) for deciding if a machine can think [6]. Actually, since the question *can a machine think?* is ill-posed and depends on the definition of thinking, Turing proposed to replace it with a practical test. The idea is that of a game in which an interrogator is given the task of determining which of two players A and B is a computer and which is a human (by using only responses to written questions). In 1956, the mathematician John McCarthy organized what remained as the famous Dartmouth workshop and which united ten prominent scientists for 2 months (among which were Marvin Minsky, Claude Shannon, Arthur Samuel, and others). This workshop is more important by its scientific program than by its outputs. Let us reproduce here the first sentences of the proposal written by McCarthy et al. [7] as we believe that they are particularly enlightening on the prospects of artificial intelligence:

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

There was no major advance made at the workshop, although a reasoning program, able to prove theorems, was presented by Allen Newell and Herbert Simon [8] at this occasion. This can be considered as the start of symbolic AI (we will come back later on the two main families of AI: symbolic and connexionist). Let us end the 1950s with the invention, in 1958, of the perceptron by Frank Rosenblatt [9], whose work was built upon the ideas of McCulloch, Pitts, and Hebb. The perceptron was the first actual artificial

neuron. It was able to recognize images. This is an important landmark for several reasons. The perceptron, with some modifications, is still the building block of modern deep learning algorithms. To mimic an artificial neuron (Fig. 2), it is composed of a set of inputs (which correspond to the information entering the synapses) x_i , which are linearly combined and then go through a non-linear function g to produce an output y . This was an important advance at the time, but it had strong limitations, in particular its inability to discriminate patterns which are not linearly separable. More generally, in the field of AI as a whole, unreasonable promises had been made, and they were not delivered: newspapers were writing about upcoming machines that could talk, see, write, and think; the US government funded huge programs to design automatic translation programs, etc. This led to a dramatic drop in research funding and, more generally, in interest in AI. This is often referred to as the first AI winter (Fig. 3).

Even though research in AI continued, it was not before the early 1980s that real-world applications were once again considered possible. This wave was that of expert systems [10], which are a type of symbolic AI approach but with domain-specific knowledge. Expert systems led to commercial applications and to a real boom in the industry. A specific programming language, called LISP [11], became dominant for the implementation of expert systems. Companies started building LISP machines, which were dedicated computers with specific architecture tailored to execute LISP efficiently. One cannot help thinking of a parallel with current hardware dedicated to deep learning. However, once again, expectations were not met. Expert systems were very large and complex sets of rules. They were difficult to maintain and update. They also had poor performances in perception tasks such as image and speech recognition. Academic and industrial funding subsequently dropped. This was the second AI winter.

At this stage, it is probably useful to come back to the two main families of AI: symbolic and connexionist (Fig. 4). They had important links at the beginning (see, e.g., the work of McCulloch and Pitt aiming to perform logical operations using artificial neurons), but they subsequently developed separately. In short, these two families can be described as follows. The first operates on symbols through sets of logical rules. It has strong ties to the domain of predicate logic. Connexionism aims at training networks of artificial neurons. This is done through the examination of training examples. More generally, it is acceptable to put most machine learning methods within the connexionist family, even though they don't rely on artificial neuron models, because their underlying principle is also to exploit statistical similarities in the training data. For a more detailed perspective on the two families of AI, the reader can refer to the very interesting (and even entertaining!) paper of Cardon et al. [12].

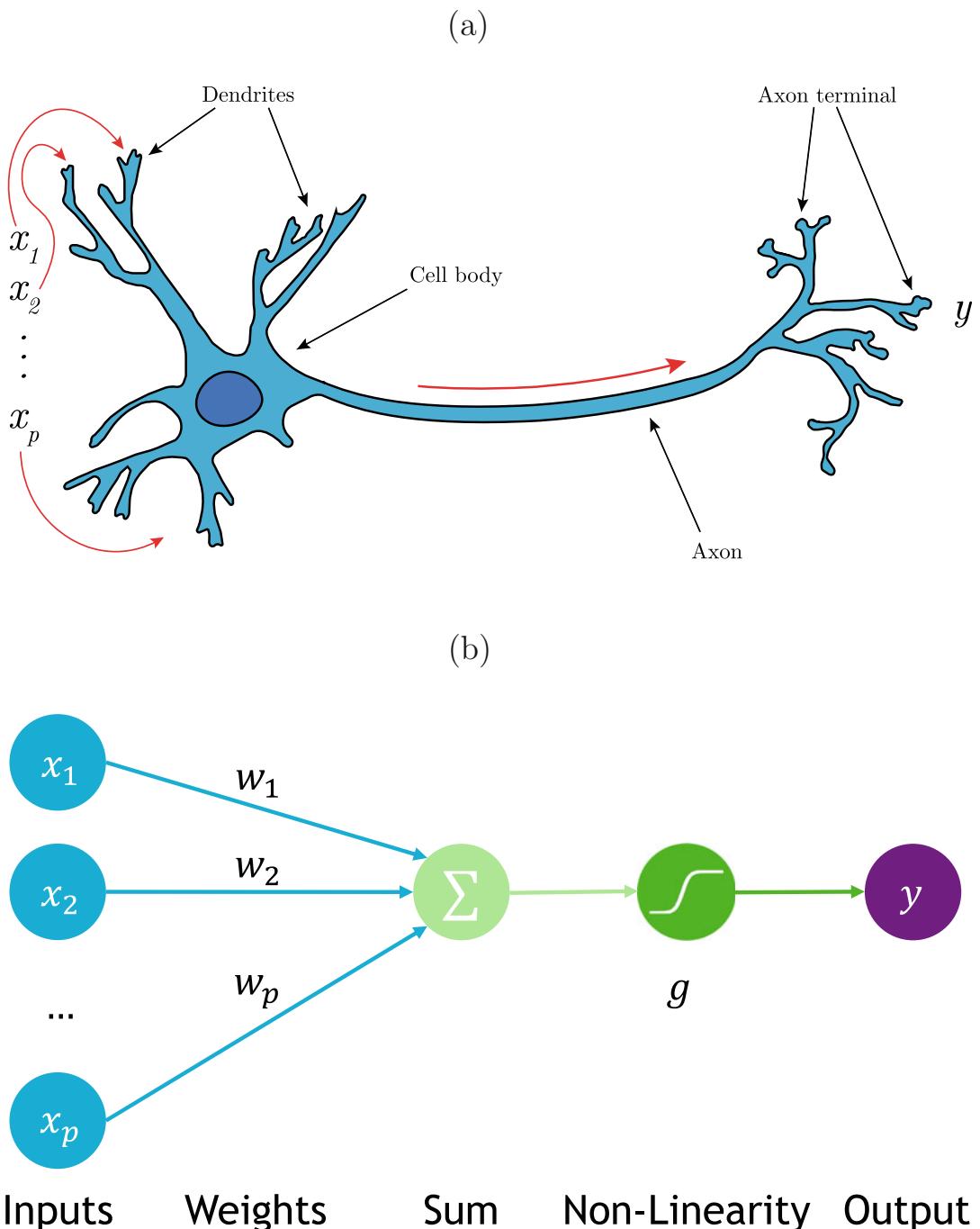


Fig. 2 (a) Biological neuron. The synapses form the input of the neuron. Their signals are combined, and if the result exceeds a given threshold, the neuron is activated and produces an output signal which is sent through the axon. (b) The perceptron: an artificial neuron which is inspired by biology. It is composed of the set of inputs (which correspond to the information entering the synapses) x_i , which are linearly combined with weights w_i and then go through a non-linear function g to produce an output y . Image in panel (a) is courtesy of Thibault Rolland

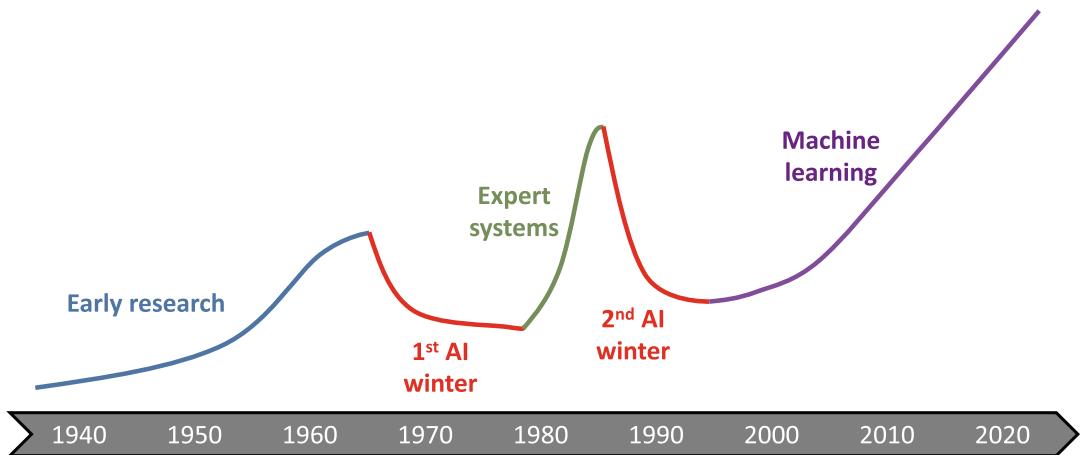


Fig. 3 Summers and winters of AI

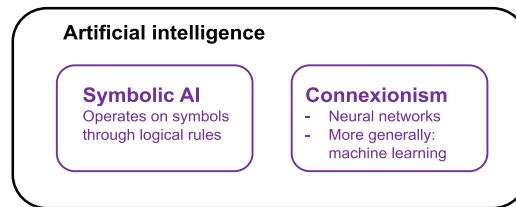


Fig. 4 Two families of AI. The symbolic approach operates on symbols through logical rules. The connexionist family actually not only encompasses artificial neural networks but more generally machine learning approaches

Let us come back to our historical timeline. The 1980s saw a rebirth of connexionism and, more generally, the start of the rise of machine learning. Interestingly, it is at that time that two of the main conferences on machine learning started: the International Conference on Machine Learning (ICML) in 1980 and Neural Information Processing Systems (NeurIPS, formerly NIPS) in 1987. It had been known for a long time that neural networks with multiple layers (as opposed to the original perceptron with a single layer) (Fig. 5) could solve non-linearly separable problems, but their training remained difficult. The back-propagation algorithm for training multilayer neural networks was described by David Rumelhart, Geoffrey Hinton, and Ronald Williams [13] in 1986, as well as by Yann LeCun in 1985 [14], who also refined the procedure in his PhD thesis published in 1987. This idea had actually been explored since the 1960s, but it was only in the 1980s that it was efficiently used for training multilayer neural networks. Finally, in 1989, Yann LeCun proposed the convolutional neural network [15], an architecture inspired by the organization of the visual cortex, whose principle is still at the core of

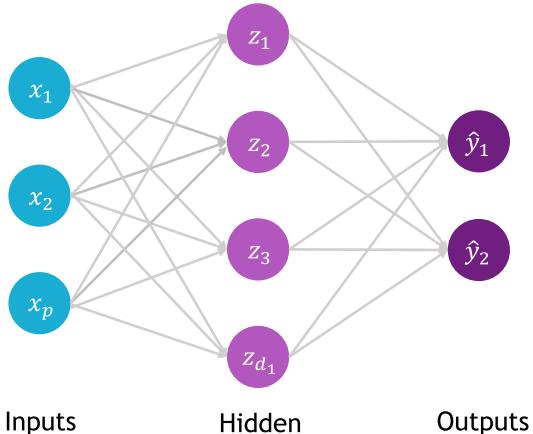


Fig. 5 A multilayer perceptron model (here with only one hidden layer, but there can be many more)

state-of-the-art algorithms for many image processing and recognition tasks. Multilayer neural networks demonstrated their utility in several real-world applications such as digit recognition on checks and ZIP codes [16]. Nevertheless, they would not become the dominant machine learning approach until the 2010s. Indeed, at the time, they required considerable computing power for training, and there was often not enough training data.

During the 1980s and 1990s, machine learning methods continued to develop. Interestingly, connections between machine learning and statistics increased. We are not going to provide an overview of the history of statistics, but one should note that many statistical methods such as linear regression [17], principal component analysis [18], discriminant analysis [19], or decision trees [20] can actually be used to solve machine learning tasks such as automatic categorization of objects or prediction. In the 1980s, decision trees witnessed important developments (see, e.g., the ID3 [21] and CART [21] algorithms). In the 1990s, there were important advances in the statistical theory of learning (in particular, the works of Vladimir Vapnik [22]). A landmark algorithm developed at that time was the support vector machine (SVM) [23] which worked well with small training datasets and could handle non-linearities through the use of kernels. The machine learning field continued to expand through the 2000s and 2010s, with new approaches but also more mature software packages such as scikit-learn [24]. More generally, it is actually important to have in mind that what is currently called AI owes more to statistics (and other mathematical fields such as optimization in particular) than to modeling of brain circuitry and that even approaches that take inspiration from neurobiology can actually be viewed as complex statistical machineries.

2012 saw the revival of neural networks and the beginning of the era of deep learning. It was undoubtedly propelled by the considerable improvement obtained on the ImageNet recognition challenge which contains 14 million natural images belonging to 20,000 categories. The solution, proposed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton [25], was a convolutional neural network with a large number of layers, hence the term deep learning. The building blocks of this solution were already present in the 1980s, but there was not enough computing power nor large training datasets for them to work properly. In the interval, things had changed. Computers had become exponentially more powerful, and, in particular, the use of graphical processing units (GPU) considerably sped up computations. The expansion of the Internet had provided massive amounts of data of various sorts such as texts and images. In the subsequent years, deep learning [26] approaches became increasingly sophisticated. In parallel, efficient and mature software packages including TensorFlow [27], PyTorch [28], or Keras [29], whose development is supported by major companies such as Google and Facebook, enable deep learning to be used more easily by scientists and engineers.

Artificial intelligence in medicine as a research field is about 50 years old. In 1975, an expert system, called MYCIN, was proposed to identify bacteria causing various infectious diseases [30]. More generally, there was a growing interest in expert systems for medical applications. Medical image processing also quickly became a growing field. The first conference on Information Processing in Medical Imaging (IPMI) was held in 1977 (it existed under a different name since 1969). The first SPIE Medical Image Processing conference took place in 1986, and the Medical Image Computing and Computer-Assisted Intervention (MICCAI) conference started in 1998. Image perception tasks, such as segmentation or classification, soon became among the key topics of this field, even though the methods came in majority from traditional image processing and not from machine learning. In the 2010s, machine learning approaches became dominant for medical image processing and more generally in artificial intelligence in medicine.

To conclude this part, it is important to be clear about the different terms, in particular those of artificial intelligence, machine learning, and deep learning (Fig. 6). Machine learning is *one* approach to artificial intelligence, and other radically different approaches exist. Deep learning is a specific type of machine learning approach. It has recently obtained impressive results on some types of data (in particular, images and text), but this does not mean that it is the universal solution to all problems. As we will see in this book, there are tasks for which other types of approaches perform best.

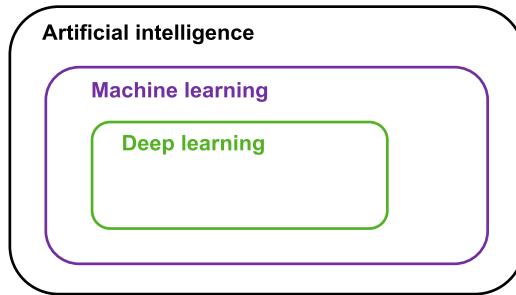


Fig. 6 Artificial intelligence, machine learning, and deep learning are not synonymous. Deep learning is a type of machine learning which involves neural networks with a large number of hidden layers. Machine learning is one approach to artificial intelligence, but other approaches exist

3 Main Machine Learning Concepts

As aforementioned, machine learning aims at making a computer capable of performing a task without explicitly being programmed for that task. More precisely, it means that one will not write a sequence of instructions that will directly perform the considered task. Instead, one will write a program that allows the computer to learn how to perform the task by examining examples or experiences. The output of this learning process is a computer program itself that performs the desired task, but this program was not explicitly written. Instead, it has been learned automatically by the computer.

In 1997, Tom Mitchell gave a more precise definition of a **well-posed machine learning problem** [31]:

A computer program is said to learn from **experience E** with respect to some **task T** and some **performance measure P**, if its performance at task T, as measured by P, improves with experience E.

He then provides the example of a computer that learns to play checkers: task T is playing checkers, performance measure P is the proportion of games won, and the training experience E is playing checker games against itself. Very often, the experience E will not be an actual action but the observation of a set of examples, for instance, a set of images belonging to different categories, such as photographs of cats and dogs, or medical images containing tumors or without lesions. Please refer to Box 1 for a summary.

Box 1: Definition of machine learning
Machine learning definition [31]:

a computer program is said to learn from **experience E** with respect to some **task T** and some **performance measure P**, if its performance at task T, as measured by P, improves with experience E.

(continued)

Box 1 (continued)

Example: learning to detect tumors from medical images

- **Task T:** detect tumors from medical image
- **Performance measure P:** proportion of tumors correctly identified
- **Experience E:** examining a dataset of medical images where the presence of tumors has been annotated

3.1 Types of Learning

One usually considers three main types of learning: supervised learning, unsupervised learning, and reinforcement learning (Box 2). In both supervised and unsupervised learning, the experience E is actually the inspection of a set of examples, which we will refer to as *training examples* or *training set*.

Box 2: Supervised, Unsupervised, and Reinforcement learning

- **Supervised learning.** Learns from labeled examples, i.e., examples for which the output that we are trying to learn is known
 - Example 1. The task is computer-aided diagnosis (a classification problem), and the label can be the diagnosis of each patient, as defined by an expert physician.
 - Example 2. The task is the prediction of the age of a person from a set of biological variables (e.g., a brain MRI). This is a regression problem. The label is the true age of a given person in the training set.
- **Unsupervised learning.** Learns from unlabeled examples
 - Example 1. Given a large set of newspaper articles, automatically cluster them into groups dealing with the same topic based only on the text of the article. The topics can, for example, be economics, politics, or international affairs. The topics are not known a priori.
 - Example 2. Given a set of patients with autism spectrum disorders, the aim is to discover a cluster of patients that share the same characteristics. The clusters are not known a priori. Examples 1 and 2 will be referred to as clustering tasks.
 - Example 3. Given a large set of medical characteristics (various biological measurements, clinical and cognitive tests, medical images), find a small set of variables that best explain the variability of the dataset. This is a dimensionality reduction problem.

(continued)

Box 2 (continued)

- **Reinforcement learning.** Learns by iteratively performing actions to maximize some reward
 - Classical approach used for learning to play games (chess, go, etc.) or in the domain of robotics
 - Currently few applications in the domain of brain diseases

3.1.1 Supervised Learning

In supervised learning, the machine learns to perform a task by examining a set of examples for which the output is known (i.e., the examples have been labeled). The two most common tasks in supervised learning are classification and regression (Fig. 7). Classification aims at assigning a category for each sample. The examples can, for instance, be different patients, and the categories are the different possible diagnoses. The outputs are thus discrete. Examples of common classification algorithms include logistic regression (in spite of its name, it is a classification method), linear discriminant analysis, support vector machines, random forest classifiers, and deep learning models for classification. In regression, the output is a continuous number. This can be, for example, the future clinical score of a patient that we are trying to predict. Examples of common regression methods include simple or multiple linear regression, penalized regression, and random forest regression. Finally, there are many other tasks that can be framed as a supervised learning problem, including, for example, data synthesis, image segmentation, and many others which will be described in other chapters of this book.

3.1.2 Unsupervised Learning

In unsupervised learning, the examples are not labeled. The two most common tasks in unsupervised learning are clustering and dimensionality reduction (Fig. 8). Clustering aims at discovering groups within the training set, but these groups are not known a priori. The objective is to find groups such that members of the same group are similar, while members of different groups are dissimilar. For example, one can aim to discover disease subtypes which are not known a priori. Some classical clustering methods are k -means or spectral clustering, for instance. Dimensionality reduction aims at finding a space of variables (of lower dimension than the input space) that best explain the variability of the training data, given a larger set of input variables. This produces a new set of variables that, in general, are not among the input variables but are combinations of them. Examples of such methods include principal component analysis, Laplacian eigenmaps, or variational autoencoders.

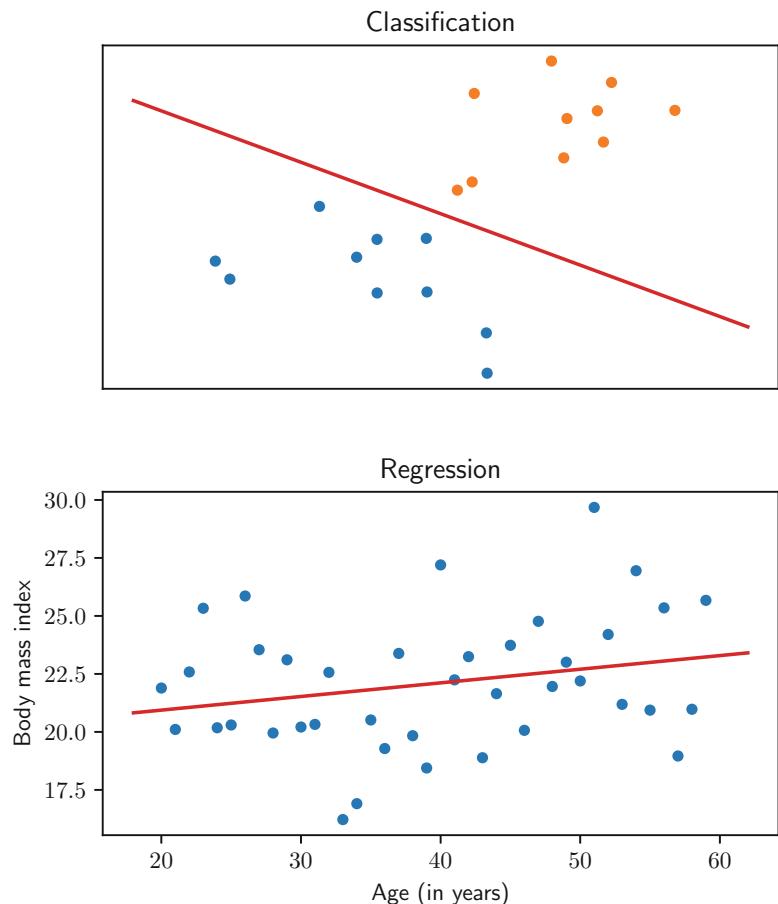


Fig. 7 Two of the main supervised learning tasks: classification and regression. The upper panel presents a classification task which aims at linearly separating the orange and the blue class. Each sample is described by two variables. The lower panel presents a linear regression task in which the aim is to predict the body mass index from the age of a person. *Figure courtesy of Johann Faouzi*

3.1.3 Reinforcement Learning

In reinforcement learning, the machine will take a series of actions in order to maximize a reward. This can, for example, be the case of a machine learning to play chess, which will play games against itself in order to maximize the number of victories. These methods are widely used for learning to play games or in the domain of robotics. So far, they have had few applications to brain diseases and will not be covered in the rest of this book.

3.1.4 Discussion

Unsupervised learning is obviously attractive because it does not require labels. Indeed, acquiring labels for a training set is usually time-consuming and expensive because the labels need to be assigned by a human. This is even more problematic in medicine because the labels must be provided by experts in the field. It is thus in principle attractive to adopt unsupervised strategies, even for

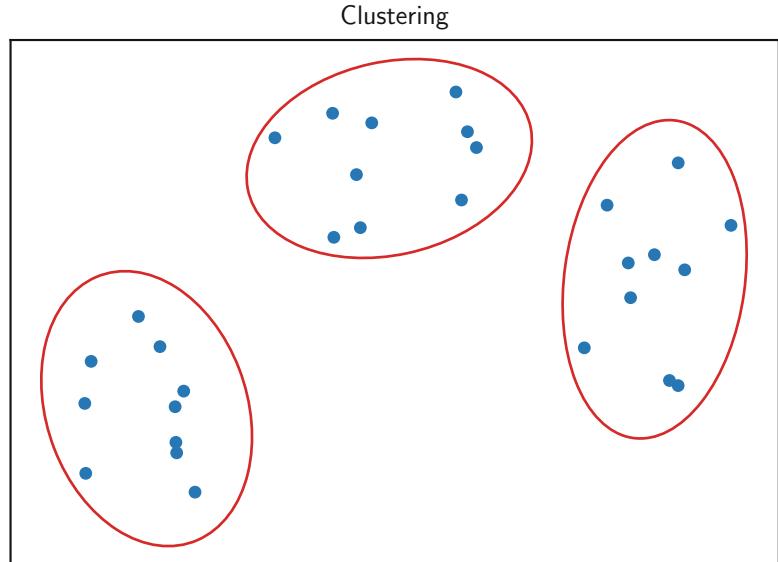


Fig. 8 Clustering task. The algorithm automatically identifies three groups (corresponding to the red circles) from unlabeled examples (the blue dots). The groups are not known a priori. *Figure courtesy of Johann Faouzi*

tasks which could be framed as supervised learning problems. Nevertheless, up to now, the performances of supervised approaches are often vastly superior in many applications. However, in the past years, an alternative strategy called self-supervised learning, where the machine itself provides its own supervision, has emerged. This is a promising approach which has already led to impressive results in different fields such as natural language processing in particular [32–34].

3.2 Overview of the Learning Process

In this section, we aim at formalizing the main concepts underlying most supervised learning methods. Some of these concepts, with modifications, also extend to unsupervised cases.

The task that we will consider will be to provide an output, denoted as y , from an input given to the computer, denoted as x . At this moment, the nature of x does not matter. It can, for example, be any possible photograph as in the example presented in Fig. 9. It could also be a single number, a series of numbers, a text, etc. For now, the nature of y can also be varied. Typically, in the case of regression, it can be a number. In the case of classification, it corresponds to a label (for instance, the label “cat” in our example). For now, you do not need to bother about how these data (images, labels, etc.) are represented in a computer. For those without a background in computer science, this will be briefly covered in Subheading 3.3.

Learning will aim at finding a function f that can transform x into y , that is, such that $y=f(x)$. For now, f can be of any type—

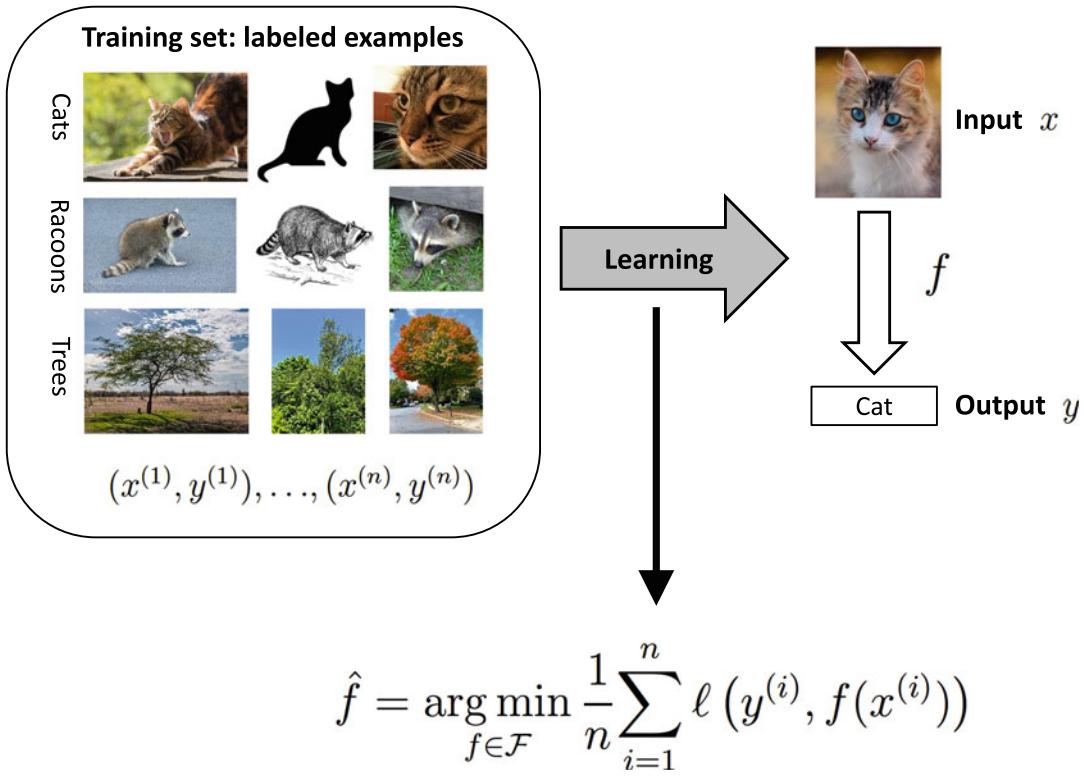


Fig. 9 Main concepts underlying supervised learning, here in the case of classification. The aim is to be able to recognize the content of a photograph (the input x) which amounts to assigning it a label (the output y). In other words, we would like to have a function f that transforms x into y . In order to find the function f , we will make use of a training set $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$ (which in our case is a set of photographs which have been labeled). All images come from <https://commons.wikimedia.org/> and have no usage restriction

just imagine it as an operation that can associate a given x with a given y . In Chap. 3, the functions f will be artificial neural networks. Learning aims at finding a function f which will provide the correct output for each given input. Let us call the loss function and denote ℓ a function that measures the error that is made by the function f . The loss function takes two arguments: the true output y and the predicted output $f(x)$. The lower the loss function value, the closer the predicted output is to the true output. An example of loss function is the classical least squares loss $\ell(y, f(x)) = (y - f(x))^2$, but many others exist. Ideally, the best function f would be the one that produces the minimal error for any possible input x and associated output y , not only those which we have at our disposal, but any other possible new data. Of course, we do not have any possible data at our disposal. Thus, we are going to use a set of data called the training set. In supervised learning, this set is labeled, i.e., for each example in this set, we know the value of both x and y . Let us denote as $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$ the n examples of the training

set which are n pairs of inputs and outputs. We are now going to search for the function f that makes the minimum error over the n samples of the training set. In other words, we are looking for the function which minimizes the average error over the training set. Let us call this average error the cost function:

$$J(f) = \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)}, f(x^{(i)}))$$

Learning will then aim at finding the function \hat{f} which minimizes the cost function:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)}, f(x^{(i)}))$$

In the above equation, *argmin* indicates that we are interested in the function f that minimizes the cost $J(f)$ and not in the value of the cost itself. \mathcal{F} is the space that contains all admissible functions. \mathcal{F} can, for instance, be the set of linear functions or the set of neural networks with a given architecture.

The procedure that will aim at finding f that minimizes the cost is called an *optimization procedure*. Sometimes, the minimum can be found analytically (i.e., by directly solving an equation for f), but this will rarely be the case. In other cases, one will resort to an iterative procedure (i.e., an algorithm): the function f is iteratively modified until we find the function which minimizes the cost. There are cases where we will have an algorithm that is guaranteed to find the global minimum and others where one will only find a local minimum.

Minimizing the errors on the training set does not guarantee that the trained computer will perform well on new examples which were not part of the training set. A first reason may be that the training set is too different from the general population (for instance, we have trained a model on a dataset of young males, and we would like to apply it to patients of any gender and age). Another reason is that, even if the training set characteristics follow those of the general population, the learned function f may be too specific to the training set. In other words, it has learned the training set “by heart” but has not discovered a more general rule that would work for other examples. This phenomenon is called *overfitting* and often arises when the dimensionality of the data is too high (there are many variables to represent an input), when the training set is too small, or when the function f is too flexible. A way to prevent overfitting will be to modify the cost function so that it not only represents the average error across training samples but also constrains the function f to have some specific properties.

Table 1
Example where the input is a series of number. Here each patient is characterized by several variables

	Age (years)	Height (cm)	Weight (kg)
Patient 1	52.5	172	52
Patient 2	75.1	182	78
Patient 3	32.7	161	47
Patient 4	45	190	92

3.3 Inputs and Features

In the previous section, we made no assumption on the nature of the input x . It could be an image, a number, a text, etc.

The simplest form of input that one can consider is when x is a single number. Examples include age, clinical scores, etc. However, for most problems, characterization of a patient cannot be done with a single number but requires a large set of measurements (Table 1). In such a case, the input can be a series of numbers x_1, \dots, x_p which can be arranged into a vector:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

However, there are cases where the input is not a vector of numbers. This is the case when the input is a medical image, a text, or a DNA sequence, for instance. Of course, in a computer, everything is stored as numbers. An image is an array of values representing the grayscale intensity of each pixel (Fig. 10). A text is a sequence of characters which are each coded as a number. However, unlike in the example presented in Table 1, these numbers are not meaningful by themselves. For this reason, a common approach is to extract features, which will be series of numbers that meaningfully represent the input. For example, if the input is a brain magnetic resonance image (MRI), relevant features could be the volumes of different anatomical regions of the brain (this specific process is done using a technique called image segmentation which is covered in another chapter). This would result in a series of numbers that would form an input vector. The development of efficient methods for extracting meaningful features from raw data is important in machine learning. Such an approach is often called *feature engineering*. Deep learning methods allow for avoiding extracting features by providing an end-to-end approach from the raw data to the output. In some areas, this has made feature engineering less important, but there are still applications where the so-called handcrafted features are competitive with deep learning methods.

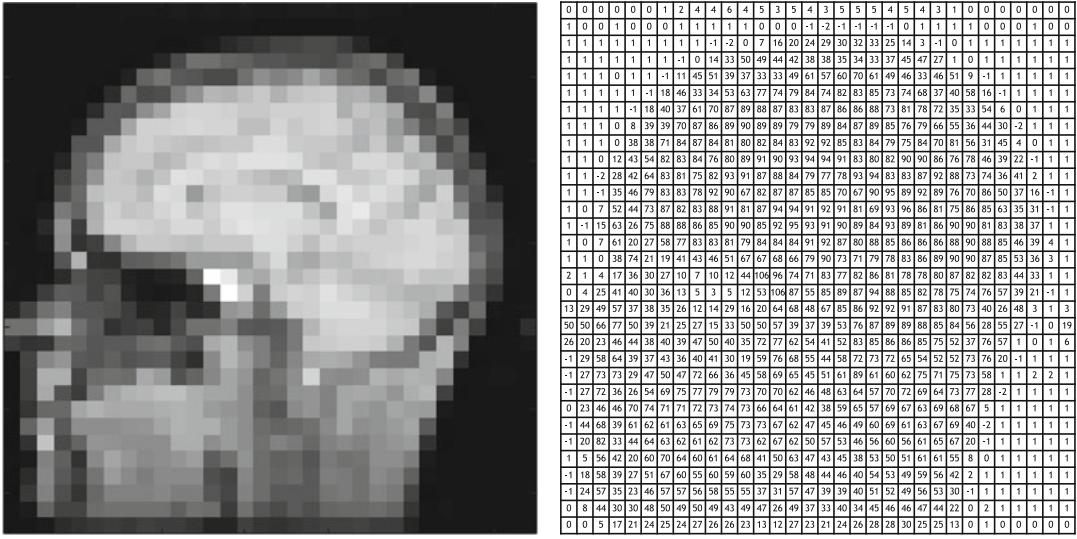


Fig. 10 In a computer, an image is represented as an array of numbers. Each number corresponds to the gray level of a given pixel. Here the example is a slice of an anatomical MRI which has been severely undersampled so that the different pixels are clearly visible. Note that an anatomical MRI is actually a 3D image and would thus be represented by a 3D array rather than by a 2D array. *Image courtesy of Ninon Burgos*

3.4 Illustration in a Simple Case

We will now illustrate step by step the above concepts in a very simple case: univariate linear regression. Univariate means that the input is a single number as in the example shown in Fig. 7. Linear means that the model f will be a simple line. The input is a number x and the output is a number y . The loss will be the least squares loss: $\ell(y, f(x)) = (y - f(x))^2$. The model f will be a linear function of x that is $f(x) = w_1 x + w_0$ and corresponds to the equation of a line, w_1 being the slope of the line and w_0 the intercept. To further simplify things, we will consider the case where there is no intercept, i.e., the line passes through the origin. Different values of w_1 correspond to different lines (and thus to different functions f) and to different values of the cost function $J(f)$, which can be in our case rewritten as $J(w_1)$ since f only depends on the parameter w_1 (Fig. 11). The best model is the one for which $J(w_1)$ is minimal.

How can we find w_1 such that $J(w_1)$ is minimal? We are going to use the derivative of J : $\frac{dJ}{dw_1}$. A minimum of $J(w_1)$ is necessarily such that $\frac{dJ}{dw_1} = 0$ (in our specific case, the converse is also true). In our case, it is possible to directly solve $\frac{dJ}{dw_1} = 0$. This will nevertheless not be the case in general. Very often, it will not be possible to solve this analytically. We will thus resort to an iterative algorithm. One classical iterative method is *gradient descent*. In the general case, f depends not on only one parameter w_1 but on a set of parameters (w_1, \dots, w_p) which can be assembled into a vector w . Thus, instead of working with the derivative $\frac{dJ}{dw_1}$, we will work with the gradient $\nabla_w J$. The gradient is a vector that indicates the direction that one should follow to climb along J . We will thus follow the opposite of the gradient, hence the name gradient descent. This process is illustrated in Fig. 12, together with the corresponding algorithm.

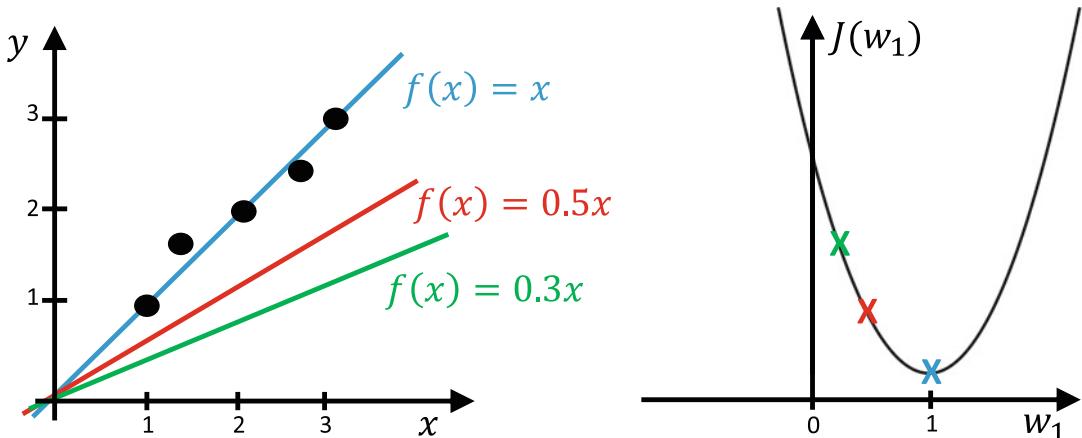


Fig. 11 We illustrate the concepts of supervised learning on a very simple case: univariate linear regression with no intercept. Training samples correspond to the black circles. The different models $f(x) = w_1 x$ correspond to the different lines. Each model (and thus each value of the parameter w_1) corresponds to a value of the cost $J(w_1)$. The best model (the blue line) is the one which minimizes $J(w_1)$; here it corresponds to the line with a slope $w_1 = 1$

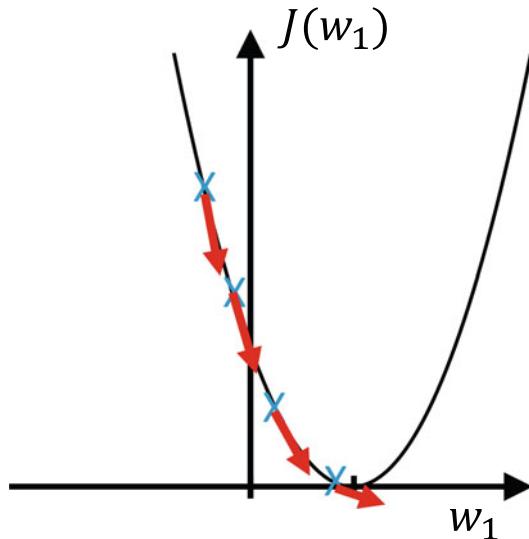


Fig. 12 Upper panel: Illustration of the concept of gradient descent in a simple case where the model f is defined using only one parameter w_1 . The value of w_1 is iteratively updated by following the opposite of the gradient. Lower panel: Gradient descent algorithm where η is the learning rate, i.e., the speed at which w_1 will be updated

4 Conclusion

This chapter provided an introduction to machine learning (ML) for a non-technical readership (e.g., physicians, neuroscientists, etc.). ML is an approach to artificial intelligence and thus needs to be put into this larger context. We introduced the main concepts underlying ML that will be further expanded in Chaps. 2–6. The reader can find a summary of these main concepts, as well as notations, in Box 3.

Box 3: Summary of main concepts

- The input x
- The output y
- The training samples $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$
- The model: transforms the input into the output
 f such that $y = f(x)$
- The set of possible models \mathcal{F}
- The loss: measures the error between the predicted and the true output, for a given sample
 $\ell(y, f(x))$
- The cost function: measures the average error across the training samples
 $J(f) = \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)}, f(x^{(i)}))$
- Learning process: finding the model which minimizes the cost function
 $\hat{f} = \arg \min_{f \in \mathcal{F}} J(f)$

Acknowledgements

The author would like to thank Johann Faouzi for his insightful comments. This work was supported by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Institut Hospitalo-Universitaire ICM).

References

1. Samuel AL (1959) Some studies in machine learning using the game of checkers. *IBM J Res Dev* 3(3):210–229
2. Russell S, Norvig P (2002) Artificial intelligence: a modern approach. Pearson, London
3. McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 5(4):115–133
4. Wiener N (1948) Cybernetics or control and communication in the animal and the machine. MIT Press, Cambridge
5. Hebb DO (1949) The organization of behavior. Wiley, New York
6. Turing AM (1950) Computing machinery and intelligence. *Mind* 59(236):433–360
7. McCarthy J, Minsky ML, Rochester N, Shannon CE (1955) A proposal for the Dartmouth summer research project on artificial intelligence. *Research Report*. <http://ravsolomonoff.com/dartmouth/boxa/dart564props.pdf>
8. Newell A, Simon H (1956) The logic theory machine—a complex information processing system. *IRE Trans Inf Theory* 2(3):61–79
9. Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 65(6):386
10. Buchanan BG, Shortliffe EH (1984) Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley, Boston
11. McCarthy J (1960) Recursive functions of symbolic expressions and their computation by machine, part I. *Commun ACM* 3(4):184–195
12. Cardon D, Cointet JP, Mazières A, Libbrecht E (2018) Neurons spike back. *Réseaux* 5:173–220. https://neurovenge.antonomase.fr/RevengeNeurons_Reseaux.pdf
13. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088):533–536
14. Le Cun Y (1985) Une procédure d'apprentissage pour réseau à seuil assymétrique. *Cognition* 85:599–604
15. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. *Neural Comput* 1(4):541–551
16. Matan O, Baird HS, Bromley J, Burges CJC, Denker JS, Jackel LD, Le Cun Y, Pednault EPD, Satterfield WD, Stenard CE et al (1992) Reading handwritten digits: a zip code recognition system. *Computer* 25(7):59–63
17. Legendre AM (1806) Nouvelles méthodes pour la détermination des orbites des comètes. Firmin Didot
18. Pearson K (1901) On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2(11):559–572
19. Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann Eugenics* 7(2):179–188
20. Loh WY (2014) Fifty years of classification and regression trees. *Int Stat Rev* 82(3):329–348
21. Quinlan JR (1986) Induction of decision trees. *Mach Learn* 1(1):81–106
22. Vapnik V (1999) The nature of statistical learning theory. Springer, Berlin
23. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on computational learning theory*, pp 144–152
24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B et al (2011) Scikit-learn: Machine learning in python. *J Mach Learn Res* 12:2825–2830
25. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, vol 25, pp 1097–1105
26. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444
27. Abadi M, Agarwal A et al (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. <https://www.tensorflow.org/>, software available from tensorflow.org
28. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) PyTorch: An imperative style, high-performance deep learning library. In: *Advances in neural information processing systems*, vol 32, pp 8026–8037

29. Chollet F et al (2015) Keras. <https://github.com/fchollet/keras>
30. Shortliffe E (1976) Computer-based medical consultations: MYCIN. Elsevier, Amsterdam
31. Mitchell T (1997) Machine learning. McGraw Hill, New York
32. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781
33. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
34. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 2

Classic Machine Learning Methods

Johann Faouzi and Olivier Colliot

Abstract

In this chapter, we present the main classic machine learning methods. A large part of the chapter is devoted to supervised learning techniques for classification and regression, including nearest neighbor methods, linear and logistic regressions, support vector machines, and tree-based algorithms. We also describe the problem of overfitting as well as strategies to overcome it. We finally provide a brief overview of unsupervised learning methods, namely, for clustering and dimensionality reduction. The chapter does not cover neural networks and deep learning as these will be presented in Chaps. 3, 4, 5, and 6.

Key words Machine learning, Classification, Regression, Clustering, Dimensionality reduction

1 Introduction

This chapter presents the main classic machine learning (ML) methods. There is a focus on supervised learning methods for classification and regression, but we also describe some unsupervised approaches. The chapter is meant to be readable by someone with no background in machine learning. It is nevertheless necessary to have some basic notions of linear algebra, probabilities, and statistics. If this is not the case, we refer the reader to Chapters 2 and 3 of [1].

The rest of this chapter is organized as follows. Rather than grouping methods by categories (for instance, classification or regression methods), we chose to present methods by increasing order of complexity. We first provide the notations in Subheading 2. We then describe a very intuitive family of methods, that of nearest neighbors (Subheading 3). We continue with linear regression (Subheading 4) and logistic regression (Subheading 5), the latter being a classification technique. We subsequently introduce the problem of overfitting (Subheading 6) as well as strategies to mitigate it (Subheading 7). Subheading 8 describes support vector machines (SVM). Subheading 9 explains how binary classification methods can be extended to a multi-class setting. We then describe

methods which are specifically adapted to the case of normal distributions (Subheading 10). Decision trees and random forests are described in Subheading 11. We then briefly describe some unsupervised learning techniques, namely, for clustering (Subheading 12) and dimensionality reduction (Subheading 13). The chapter ends with a description of kernel methods which can be used to extend linear techniques to non-linear cases (Subheading 14). **Box 1** summarizes the methods presented in this chapter, grouped by categories and then sorted in order of appearance.

Box 1: Main Classic ML Methods

- **Supervised learning**
 - **Classification:** nearest neighbors, logistic regression, support vector machine (SVM), naive Bayes, linear discriminant analysis (LDA), quadratic discriminant analysis, tree-based models (decision tree, random forest, extremely randomized trees)
 - **Regression:** nearest neighbors, linear regression, support vector machine regression, tree-based models (decision tree, random forest, extremely randomized trees), kernel ridge regression
- **Unsupervised learning**
 - **Clustering:** k -means, Gaussian mixture model
 - **Dimensionality reduction:** principal component analysis (PCA), linear discriminant analysis (LDA), kernel principal component analysis

2 Notations

Let n be the number of samples and p be the number of features. An input sample is thus a p -dimensional vector:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

An output sample is denoted by y . Thus, a sample is (\boldsymbol{x}, y) . The dataset of n samples can then be summarized as an $n \times p$ matrix \boldsymbol{X} representing the input data and an n -dimensional vector \boldsymbol{y} representing the target data:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \vdots \\ \mathbf{x}^{(n)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & \dots & x_p^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(n)} & \dots & x_p^{(n)} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

The input space is denoted by I , and the set of training samples is denoted by \mathcal{X} .

In the case of regression, y is a real number. In the case of classification, y is a single label. More precisely, y can only take one of a finite set of values called labels. The set of possible classes (i.e., labels) is denoted by $\mathcal{C} = \{C_1, \dots, C_q\}$, with q being the number of classes. As the values of the classes are not meaningful, when there are only two classes, the classes are often called the positive and negative classes. In this case and also for mathematical reasons, without loss of generality, we assume the values of the classes to be $+1$ and -1 .

3 Nearest Neighbor Methods

One of the most intuitive approaches to machine learning is nearest neighbors. It is based on the following intuition: for a given input, its corresponding output is likely to be similar to the outputs of similar inputs. A real-life metaphor would be that if a subject has similar characteristics than other subjects who were diagnosed with a given disease, then this subject is likely to also be suffering from this disease.

More formally, nearest neighbor methods use the training samples from the neighborhood of a given point \mathbf{x} , denoted by $N(\mathbf{x})$, to perform prediction [2].

For regression tasks, the prediction is computed as a weighted mean of the target values in $N(\mathbf{x})$:

$$\hat{y} = \sum_{\mathbf{x}^{(i)} \in N(\mathbf{x})} w_i^{(\mathbf{x})} y^{(i)}$$

where $w_i^{(\mathbf{x})}$ is the weight associated with $\mathbf{x}^{(i)}$ to predict the output of \mathbf{x} , with $w_i^{(\mathbf{x})} \geq 0 \forall i$ and $\sum_i w_i^{(\mathbf{x})} = 1$.

For classification tasks, the predicted label corresponds to the label with the largest weighted sum of occurrences of each label:

$$\hat{y} = \arg \max_{\mathcal{C}} \sum_{\mathbf{x}^{(i)} \in N(\mathbf{x})} w_i^{(\mathbf{x})} \mathbf{1}_{y^{(i)} = C_k}$$

A key parameter of nearest neighbor methods is the *metric*, denoted by d , that is, a mathematical function that defines dissimilarity. The metric is used to define the neighborhood of any point and can also be used to compute the weights.

3.1 Metrics

Many metrics have been defined for various types of input data such as vectors of real numbers, integers, or booleans. Among these different types, vectors of real numbers are one of the most common types of input data, for which the most commonly used metric is the Euclidean distance, defined as:

$$\forall \mathbf{x}, \mathbf{x}' \in I, \|\mathbf{x} - \mathbf{x}'\|_2 = \sqrt{\sum_{j=1}^p (x_j - x'_j)^2}$$

The Euclidean distance is sometimes referred to as the “ordinary” distance since it is the one based on the Pythagorean theorem and that everyone uses in their everyday lives.

3.2 Neighborhood

The two most common definitions of the neighborhood rely on either the number of neighbors or the radius around the given point. Figure 1 illustrates the differences between both definitions.

The k -nearest neighbor method defines the neighborhood of a given point \mathbf{x} as the set of the k closest points to \mathbf{x} :

$$N(\mathbf{x}) = \{\mathbf{x}^{(i)}\}_{i=1}^k \quad \text{with} \quad d(\mathbf{x}, \mathbf{x}^{(1)}) \leq \dots \leq d(\mathbf{x}, \mathbf{x}^{(n)})$$

The radius neighbor method defines the neighborhood of a given point \mathbf{x} as the set of points whose dissimilarity to \mathbf{x} is smaller than the given radius, denoted by r :

$$N(\mathbf{x}) = \{\mathbf{x}^{(i)} \in \mathcal{X} \mid d(\mathbf{x}, \mathbf{x}^{(i)}) < r\}$$

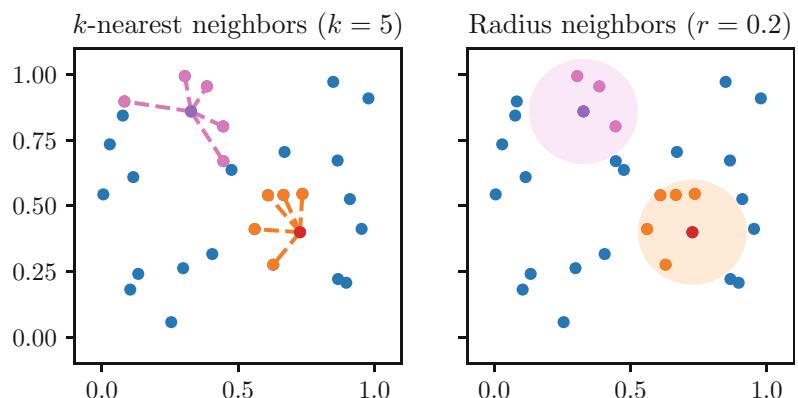


Fig. 1 Different definitions of the neighborhood. On the left, the neighborhood of a given point is the set of its five nearest neighbors. On the right, the neighborhood of a given point is the set of points whose dissimilarity is lower than the radius. For a given input, its neighborhood may be different depending on the definition used. The Euclidean distance is used as the metric in both examples

3.3 Weights

The two most common approaches to compute the weights are to use:

- Uniform weights (all the weights are equal):

$$\forall i, w_i^{(\mathbf{x})} = \frac{1}{|N(\mathbf{x})|}$$

- Weights inversely proportional to the dissimilarity:

$$\forall i, w_i^{(\mathbf{x})} = \frac{\frac{1}{d(\mathbf{x}^{(i)}, \mathbf{x})}}{\sum_j \frac{1}{d(\mathbf{x}^{(j)}, \mathbf{x})}} = \frac{1}{d(\mathbf{x}^{(i)}, \mathbf{x}) \sum_j \frac{1}{d(\mathbf{x}^{(j)}, \mathbf{x})}}$$

With uniform weights, every point in the neighborhood equally contributes to the prediction. With weights inversely proportional to the dissimilarity, closer points contribute more to the prediction than further points. Figure 2 illustrates the different decision functions obtained with uniform weights and weights inversely proportional to the dissimilarity for a 3-nearest neighbor classification model.

3.4 Neighbor Search

The brute-force method to compute the neighborhood for n points with p features is to compute the metric for each pair of inputs, which has a $\mathcal{O}(n^2 p)$ algorithmic complexity (assuming that evaluating the metric for a pair of inputs has a complexity of $\mathcal{O}(p)$, which is the case for most metrics). However, it is possible to decrease this algorithmic complexity if the metric d satisfies the following properties:

1. Non-negativity: $\forall \mathbf{a}, \mathbf{b}, d(\mathbf{a}, \mathbf{b}) \geq 0$
2. Identity: $\forall \mathbf{a}, \mathbf{b}, d(\mathbf{a}, \mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$

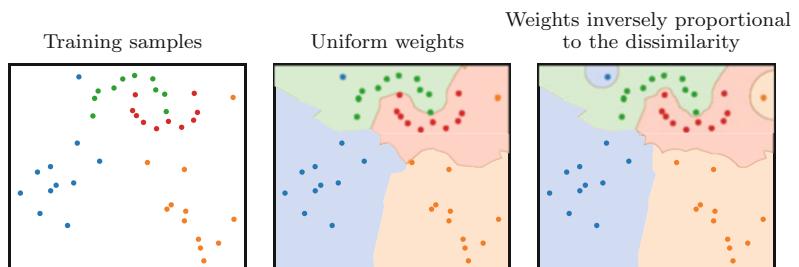


Fig. 2 Impact of the definition of the weights on the prediction function of a 3-nearest neighbor classification model. When the weights are inversely proportional to the dissimilarity, the classifier is more subject to outliers since the predictions in the close neighborhood of any input are mostly dedicated by the label of this input, independently of the number of neighbors used. With uniform weights, the prediction function tends to be smoother

3. Symmetry: $\forall \mathbf{a}, \mathbf{b}, d(\mathbf{a}, \mathbf{b}) = d(\mathbf{b}, \mathbf{a})$

4. Triangle inequality: $\forall \mathbf{a}, \mathbf{b}, \mathbf{c}, d(\mathbf{a}, \mathbf{b}) + d(\mathbf{b}, \mathbf{c}) \geq d(\mathbf{a}, \mathbf{c})$

The key property is the *triangle inequality*, which has a simple interpretation: the shortest path between two points is a straight line. Mathematically, if \mathbf{a} is far from \mathbf{c} and \mathbf{c} is close to \mathbf{b} (i.e., $d(\mathbf{a}, \mathbf{c})$ is large and $d(\mathbf{b}, \mathbf{c})$ is small), then \mathbf{a} is far from \mathbf{b} (i.e., $d(\mathbf{a}, \mathbf{b})$ is large). This is obtained by rewriting the triangle inequality as follows:

$$\forall \mathbf{a}, \mathbf{b}, \mathbf{c}, d(\mathbf{a}, \mathbf{b}) \geq d(\mathbf{a}, \mathbf{c}) - d(\mathbf{b}, \mathbf{c})$$

This means that it is not necessary to compute $d(\mathbf{a}, \mathbf{b})$ in this case. Therefore, the computational cost of a nearest neighbor search can be reduced to $\mathcal{O}(n \log(n)p)$ or better, which is a substantial improvement over the brute-force method for large n . Two popular methods that take advantage of this property are the *K-dimensional tree* structure [3] and the *ball tree* structure [4].

4 Linear Regression

Linear regression is a regression model that linearly combines the features. Each feature is associated with a coefficient that represents the relative weight of this feature compared to the other features. A real-life metaphor would be to see the coefficients as the ingredients of a recipe: the key is to find the best balance (i.e., proportions) between all the ingredients in order to make the best cake.

Mathematically, a linear model is a model that linearly combines the features [5]:

$$f(\mathbf{x}) = w_0 + \sum_{j=1}^p w_j x_j$$

A common notation consists in including a 1 in \mathbf{x} so that $f(\mathbf{x})$ can be written as the *dot product* between the vector \mathbf{x} and the vector \mathbf{w} :

$$f(\mathbf{x}) = w_0 \times 1 + \sum_{j=1}^p w_j x_j = \mathbf{x}^\top \mathbf{w}$$

where the vector \mathbf{w} consists of:

- The intercept (also known as bias) w_0
- The coefficients (w_1, \dots, w_p), where each coefficient w_j is associated with the corresponding feature x_j

In the case of linear regression, $f(\mathbf{x})$ is the predicted output:

$$\hat{y} = f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}$$

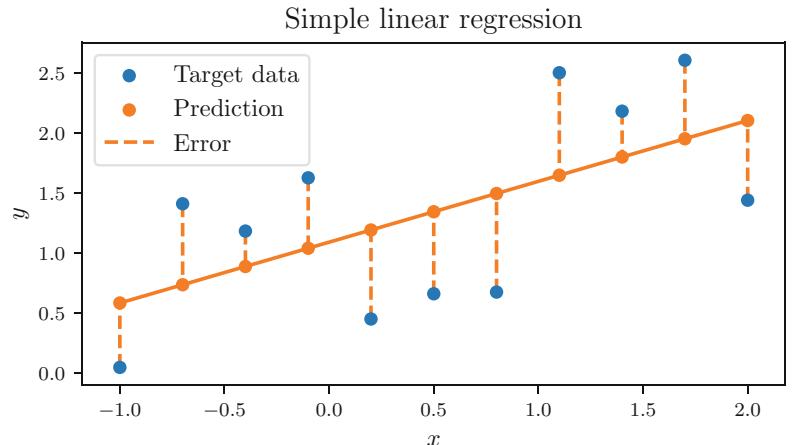


Fig. 3 Ordinary least squares regression. The coefficients (i.e., the intercept and the slope with a single predictor) are estimated by minimizing the sum of the squared errors

There are several methods to estimate the \mathbf{w} coefficients. In this section, we present the oldest one which is known as *ordinary least squares regression*.

In the case of ordinary least squares regression, the cost function J is the sum of the squared errors on the training data (see Fig. 3):

$$J(\mathbf{w}) = \sum_{i=1}^n \left(y^{(i)} - \hat{y}^{(i)} \right)^2 = \sum_{i=1}^n \left(y^{(i)} - \mathbf{x}^{(i)\top} \mathbf{w} \right)^2 = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

One wants to find the optimal parameters \mathbf{w}^* that minimize the cost function:

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} J(\mathbf{w})$$

This optimization problem is *convex*, implying that any local minimum is a global minimum, and *differentiable*, implying that every local minimum has a null gradient. One therefore aims to find null gradients of the cost function:

$$\begin{aligned} \nabla_{\mathbf{w}^*} J &= 0 \\ \Rightarrow 2\mathbf{X}^\top \mathbf{X}\mathbf{w}^* - 2\mathbf{X}^\top \mathbf{y} &= 0 \\ \Rightarrow \mathbf{X}^\top \mathbf{X}\mathbf{w}^* &= \mathbf{X}^\top \mathbf{y} \\ \Rightarrow \mathbf{w}^* &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \end{aligned}$$

Ordinary least squares regression is one of the few machine learning optimization problems for which there exists a *closed formula*, i.e., the optimal solution can be computed using a finite number of standard operations such as addition, multiplication,

and evaluations of well-known functions. A summary of linear regression can be found in Box 2.

Box 2: Linear Regression

- **Main idea:** best hyperplane (i.e., line when $p=1$, plane when $p=2$) mapping the inputs and to the outputs.
- **Mathematical formulation:** linear relationship between the predicted output \hat{y} and the input \mathbf{x} that minimizes the sum of squared errors:

$$\hat{y} = \mathbf{w}_0^* + \sum_{j=1}^n w_j^* x_j \quad \text{with} \quad \mathbf{w}^* = \arg \min_{\mathbf{w}} \sum_{i=1}^n (y^{(i)} - \mathbf{x}^{(i)\top} \mathbf{w})^2$$

- **Regularization:** can be penalized to avoid overfitting (ridge), to perform feature selection (lasso), or both (elastic-net). See Subheading 7.

5 Logistic Regression

Intuitively, linear regression consists in finding the line that best fits the data: the true output should be as close to the line as possible. For binary classification, one wants the line to separate both classes as well as possible: the samples from one class should all be in one subspace, and the samples from the other class should all be in the other subspace, with the inputs being as far as possible from the line.

Mathematically, for binary classification tasks, a linear model is defined by a hyperplane splitting the input space into two subspaces such that each subspace is characteristic of one class. For instance, a line splits a plane into two subspaces in the two-dimensional case, while a plane splits a three-dimensional space into two subspaces. A hyperplane is defined by a vector $\mathbf{w} = (w_0, w_1, \dots, w_p)$, and $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}$ corresponds to the *signed distance* between the input \mathbf{x} and the hyperplane \mathbf{w} : in one subspace, the distance with any input is always positive, whereas in the other subspace, the distance with any input is always negative. Figure 4 illustrates the decision function in the two-dimensional case where both classes are linearly separable.

The sign of the signed distance corresponds to the decision function of a linear binary classification model:

$$\hat{y} = \text{sign}(f(\mathbf{x})) = \begin{cases} +1 & \text{if } f(\mathbf{x}) > 0 \\ -1 & \text{if } f(\mathbf{x}) < 0 \end{cases}$$

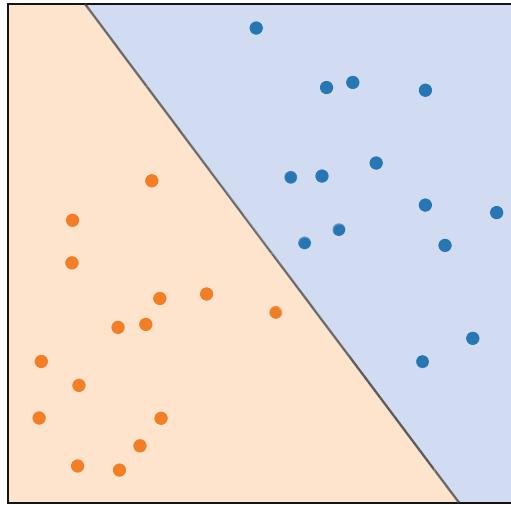


Fig. 4 Decision function of a logistic regression model. A logistic regression is a linear model, that is, its decision function is linear. In the two-dimensional case, it separates a plane with a line

The logistic regression model is a probabilistic linear model that transforms the signed distance to the hyperplane into a probability using the sigmoid function [6], denoted by $\sigma(u) = \frac{1}{1 + \exp(-u)}$.

Consider the linear model:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} = w_0 + \sum_{i=1}^p w_i x_i$$

Then the probability of belonging to the positive class is:

$$P(y = +1 | \mathbf{x} = \mathbf{x}) = \sigma(f(\mathbf{x})) = \frac{1}{1 + \exp(-f(\mathbf{x}))}$$

and that of belonging to the negative class is:

$$\begin{aligned} P(y = -1 | \mathbf{x} = \mathbf{x}) &= 1 - P(y = +1 | \mathbf{x} = \mathbf{x}) \\ &= \frac{\exp(-f(\mathbf{x}))}{1 + \exp(-f(\mathbf{x}))} \\ &= \frac{1}{1 + \exp(f(\mathbf{x}))} \\ P(y = -1 | \mathbf{x} = \mathbf{x}) &= \sigma(-f(\mathbf{x})) \end{aligned}$$

By applying the inverse of the sigmoid function, which is known as the logit function, one can see that the logarithm of the *odds ratio* is modeled as a linear combination of the features:

$$\log\left(\frac{P(y = +1 | \mathbf{x} = \mathbf{x})}{P(y = -1 | \mathbf{x} = \mathbf{x})}\right) = \log\left(\frac{P(y = +1 | \mathbf{x} = \mathbf{x})}{1 - P(y = +1 | \mathbf{x} = \mathbf{x})}\right) = f(\mathbf{x})$$

The \mathbf{w} coefficients are estimated by maximizing the *likelihood* function, that is, the function measuring the goodness of fit of the model to the training data:

$$L(\mathbf{w}) = \prod_{i=1}^n P(y=y^{(i)} | \mathbf{x}=\mathbf{x}^{(i)}; \mathbf{w})$$

For computational reasons, it is easier to maximize the *log-likelihood*, which is simply the logarithm of the likelihood:

$$\begin{aligned} \log(L(\mathbf{w})) &= \sum_{i=1}^n \log(P(y=y^{(i)} | \mathbf{x}=\mathbf{x}^{(i)}; \mathbf{w})) \\ &= \sum_{i=1}^n \log(\sigma(y^{(i)} f(\mathbf{x}^{(i)}; \mathbf{w}))) \\ &= \sum_{i=1}^n -\log(1 + \exp(y^{(i)} \mathbf{x}^{(i)\top} \mathbf{w})) \\ \log(L(\mathbf{w})) &= -\sum_{i=1}^n \log(1 + \exp(y^{(i)} \mathbf{x}^{(i)\top} \mathbf{w})) \end{aligned}$$

Finally, we can rewrite this maximization problem as a minimization problem by noticing that $\max_{\mathbf{w}} \log(L(\mathbf{w})) = -\min_{\mathbf{w}} -\log(L(\mathbf{w}))$:

$$\max_{\mathbf{w}} \log(L(\mathbf{w})) = -\min_{\mathbf{w}} \sum_{i=1}^n \log(1 + \exp(y^{(i)} \mathbf{x}^{(i)\top} \mathbf{w}))$$

We can see that the \mathbf{w} coefficients that maximize the likelihood are also the coefficients that minimize the sum of the *logistic loss* values, with the logistic loss being defined as:

$$\ell_{\text{logistic}}(y, f(\mathbf{x})) = \log(1 + \exp(yf(\mathbf{x}))) / \log(2)$$

Unlike for linear regression, there is no closed formula for this minimization. One thus needs to use an optimization method such as gradient descent which was presented in Subheading 3 of Chap. 1. In practice, more sophisticated approaches such as quasi-Newton methods and variants of stochastic gradient descent are often used. The main concepts underlying logistic regression can be found in Box 3.

Box 3: Logistic Regression

- **Main idea:** best hyperplane (i.e., line) that separates two classes.
- **Mathematical formulation:** the signed distance to the hyperplane is mapped into the probability to belong to the positive class using the sigmoid function:

(continued)

Box 3 (continued)

$$f(\mathbf{x}) = w_0 + \sum_{j=1}^n w_j x_j$$

$$P(y = +1 | \mathbf{x} = \mathbf{x}) = \sigma(f(\mathbf{x})) = \frac{1}{1 + \exp(-f(\mathbf{x}))}$$

- **Estimation:** likelihood maximization.
- **Regularization:** can be penalized to avoid overfitting (ℓ_2 penalty), to perform feature selection (ℓ_1 penalty), or both (elastic-net penalty).

6 Overfitting and Regularization

The original formulations of ordinary least squares regression and logistic regression are *unregularized* models, that is, the model is trained to fit the training data as much as possible. Let us consider a real-life example as it is very similar to human learning. If a person learns by heart the content of a book, they are able to solve the exercises in the book, but unable to apply the theoretical concepts to new exercises or real-life situations. If a person only quickly reads through the book, they are probably unable to solve neither the exercises in the book nor new exercises.

The corresponding concepts are known as *overfitting* and *underfitting* in machine learning. Overfitting occurs when a model fits too well the training data and generalizes poorly to new data. Oppositely, underfitting occurs when a model does not capture well enough the characteristics of the training data and thus also generalizes poorly to new data.

Overfitting and underfitting are related to frequently used terms in machine learning: *bias* and *variance*. Bias is defined as the expected (i.e., mean) difference between the true output and the predicted output. Variance is defined as the variability of the predicted output. For instance, let us consider a model predicting the age of a person from a picture. If the model *always* underestimates or overestimates the age, then the model is biased. If the model makes *both large and small errors*, then the model has a high variance.

Ideally, one would like to have a model with a small bias and a small variance. However, the bias of a model tends to increase when decreasing its variance, and the variance of the model tends to increase when decreasing its bias. This phenomenon is known as the *bias-variance trade-off*. Figure 5 illustrates this phenomenon. One can also notice it by computing the squared error between the true output y (fixed) and the predicted output \hat{y} (random variable): its expected value is the sum of the squared bias of \hat{y} and the variance of \hat{y} :

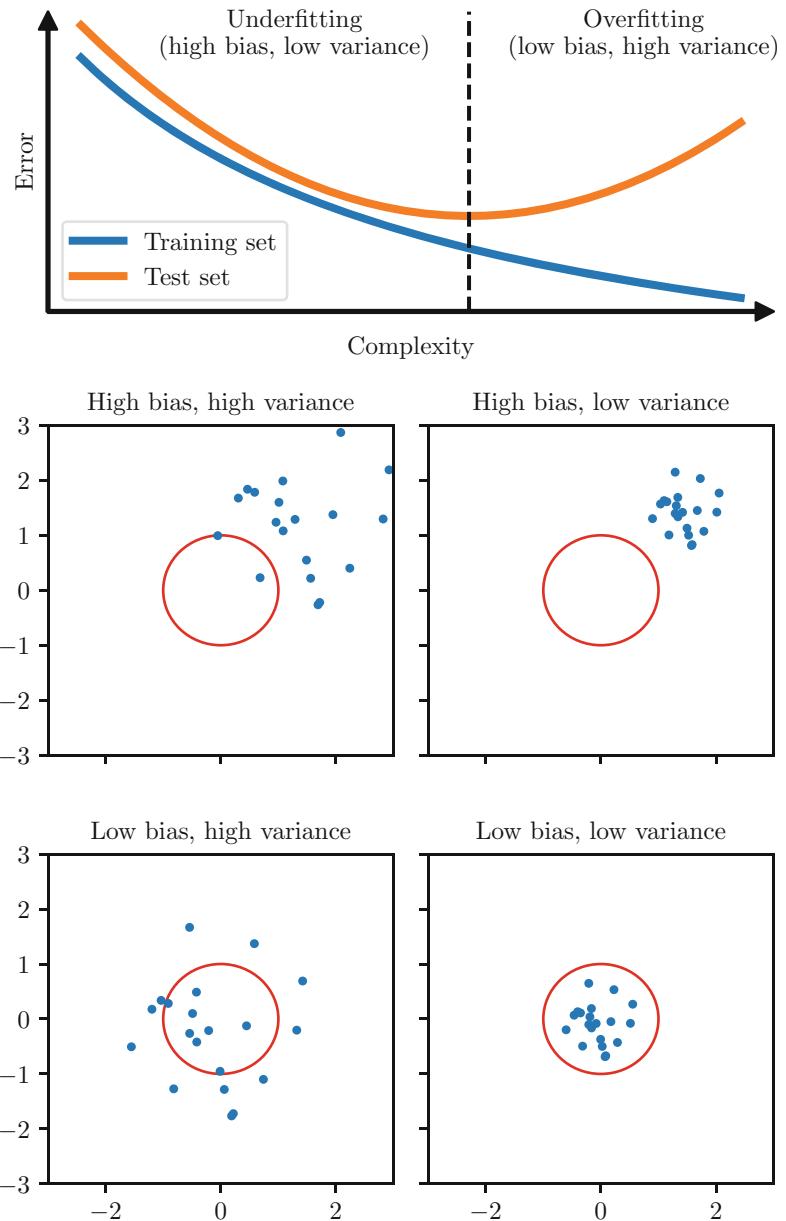


Fig. 5 Illustration of underfitting and overfitting. Underfitting occurs when a model is too simple and does not capture well enough the characteristics of the training data, leading to high bias and low variance. Oppositely, overfitting occurs when a model is too complex and learns the noise in the training data, leading to low bias and high variance

$$\begin{aligned}
\mathbb{E}[(y - \hat{y})^2] &= \mathbb{E}[y^2 - 2y\hat{y} + \hat{y}^2] \\
&= y^2 - 2y\mathbb{E}[\hat{y}] + \mathbb{E}[\hat{y}^2] \\
&= y^2 - 2y\mathbb{E}[\hat{y}] + \mathbb{E}[\hat{y}^2] + \mathbb{E}[\hat{y}]^2 - \mathbb{E}[\hat{y}]^2 \\
&= (\mathbb{E}[\hat{y}] - y)^2 + \mathbb{E}[\hat{y}^2] - \mathbb{E}[\hat{y}]^2 \\
&= (\mathbb{E}[\hat{y}] - y)^2 + \mathbb{E}[\hat{y}^2 - \mathbb{E}[\hat{y}]^2] \\
&= (\mathbb{E}[\hat{y}] - y)^2 + \mathbb{E}[\hat{y}^2 - 2\mathbb{E}[\hat{y}]^2 + \mathbb{E}[\hat{y}]^2] \\
&= (\mathbb{E}[\hat{y}] - y)^2 + \mathbb{E}[\hat{y}^2 - 2\hat{y}\mathbb{E}[\hat{y}] + \mathbb{E}[\hat{y}]^2] \\
&= (\mathbb{E}[\hat{y}] - y)^2 + \mathbb{E}[(\hat{y} - \mathbb{E}[\hat{y}])^2]
\end{aligned}$$

$$\mathbb{E}[(y - \hat{y})^2] = \underbrace{(\mathbb{E}[\hat{y}] - y)^2}_{\text{bias}^2} + \underbrace{\text{Var}[\hat{y}]}_{\text{variance}}$$

7 Penalized Models

Depending on the class of methods, there exist different strategies to tackle overfitting.

For neighbor methods, the number of neighbors used to define the neighborhood of any input and the strategy to compute the weights are the key hyperparameters to control the bias-variance trade-off. For models that are presented in the remaining sections of this chapter, we mention strategies to address the bias-variance trade-off in their respective sections. In this section, we present the most commonly used strategies for models whose parameters are optimized by minimizing a cost function defined as the mean loss values over all the training samples:

$$\min_{\boldsymbol{w}} J(\boldsymbol{w}) \quad \text{with} \quad J(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)}, f(\boldsymbol{x}^{(i)}; \boldsymbol{w}))$$

This is, for instance, the case of the linear and logistic regression methods presented in the previous sections.

7.1 Penalties

The main idea is to introduce a *penalty term* $\text{Pen}(\boldsymbol{w})$ that will constraint the parameters \boldsymbol{w} to have some desired properties. The most common penalties are the ℓ_2 penalty, the ℓ_1 penalty, and the elastic-net penalty.

7.1.1 ℓ_2 Penalty

The ℓ_2 penalty is defined as the squared ℓ_2 norm of the \boldsymbol{w} coefficients:

$$\ell_2(\mathbf{w}) = \|\mathbf{w}\|_2^2 = \sum_{j=1}^p w_j^2$$

The ℓ_2 penalty forces each coefficient w_i not to be too large and makes the coefficients more robust to collinearity (i.e., when some features are approximately linear combinations of the other features).

7.1.2 ℓ_1 Penalty

The ℓ_2 penalty forces the values of the parameters not to be too large, but does not incentivize to make small values tend to zero. Indeed, the square of a small value is even smaller. When the number of features is large, or when interpretability is important, it can be useful to make the model select the most important features. The corresponding metric is the ℓ_0 “norm” (which is not a proper norm in the mathematical sense), defined as the number of nonzero elements:

$$\ell_0(\mathbf{w}) = \|\mathbf{w}\|_0 = \sum_{j=1}^p \mathbf{1}_{w_j \neq 0}$$

However, the ℓ_0 “norm” is neither differentiable nor convex (which are useful properties to solve an optimization problem, but this is not further detailed for the sake of conciseness). The best convex differentiable approximation of the ℓ_0 “norm” is the ℓ_1 norm (see Fig. 6), defined as the sum of the absolute values of each element:

$$\ell_1(\mathbf{w}) = \|\mathbf{w}\|_1 = \sum_{j=1}^p |w_j|$$

7.1.3 Elastic-Net Penalty

Both the ℓ_2 and ℓ_1 penalties have their upsides and downsides. In order to try to obtain the best of penalties, one can add both penalties in the objective function. The combination of both penalties is known as the *elastic-net* penalty:

$$\text{EN}(\mathbf{w}, \alpha) = \alpha \|\mathbf{w}\|_1 + (1 - \alpha) \|\mathbf{w}\|_2^2$$

where $\alpha \in [0, 1]$ is a hyperparameter representing the proportion of the ℓ_1 penalty compared to the ℓ_2 penalty.

7.2 New Optimization Problem

A natural approach would be to add a constraint to the minimization problem:

$$\min_{\mathbf{w}} J(\mathbf{w}) \quad \text{subject to} \quad \text{Pen}(\mathbf{w}) < c \quad (1)$$

which reads as “Find the optimal parameters that minimize the cost function J among all the parameters \mathbf{w} that satisfy $\text{Pen}(\mathbf{w}) < c$ ” for a positive real number c . Figure 7 illustrates the optimal solution of a simple linear regression task with different constraints. This figure

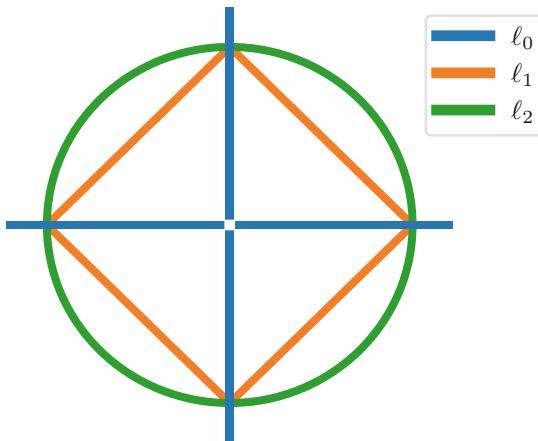


Fig. 6 Unit balls of the ℓ_0 , ℓ_1 , and ℓ_2 norms. For each norm, the set of points in \mathbb{R}^2 whose norm is equal to 1 is plotted. The ℓ_1 norm is the best convex approximation to the ℓ_0 norm. Note that the lines for the ℓ_0 norm extend to $-\infty$ and $+\infty$ but are cut for plotting reasons

also highlights the sparsity property of the ℓ_1 penalty (the optimal parameter for the horizontal axis is set to zero) that the ℓ_2 penalty does not have (the optimal parameter for the horizontal axis is small but different from zero).

Although this approach is appealing due to its intuitiveness and the possibility to set the maximum possible penalty on the parameters \mathbf{w} , it leads to a minimization problem that is not trivial to solve. A similar approach consists in adding the regularization term in the cost function:

$$\min_{\mathbf{w}} J(\mathbf{w}) + \lambda \times \text{Pen}(\mathbf{w}) \quad (2)$$

where $\lambda > 0$ is a hyperparameter that controls the weights of the penalty term compared to the mean loss values over all the training samples. This formulation is related to the Lagrangian function of the minimization problem with the penalty constraint.

This formulation leads to a minimization problem with no constraint which is much easier to solve. One can actually show that Eqs. 1 and 2 are related: solving Eq. 2 for a given λ , whose optimal solution is denoted by \mathbf{w}_λ^* , is equivalent to solving Eq. 1 for $c = \text{Pen}(\mathbf{w}_\lambda^*)$. In other words, solving Eq. 2 for a given λ is equivalent to solving Eq. 1 for c whose value is only known after finding the optimal solution of Eq. 2.

Figure 8 illustrates the impact of the regularization term $\lambda \times \text{Pen}(\mathbf{w})$ on the prediction function of a kernel ridge regression algorithm (*see* Subheading 14 for more details) for different values of λ . For high values of λ , the regularization term is dominating the mean loss value, making the prediction function not fitting well enough the training data (underfitting). For small values of λ , the

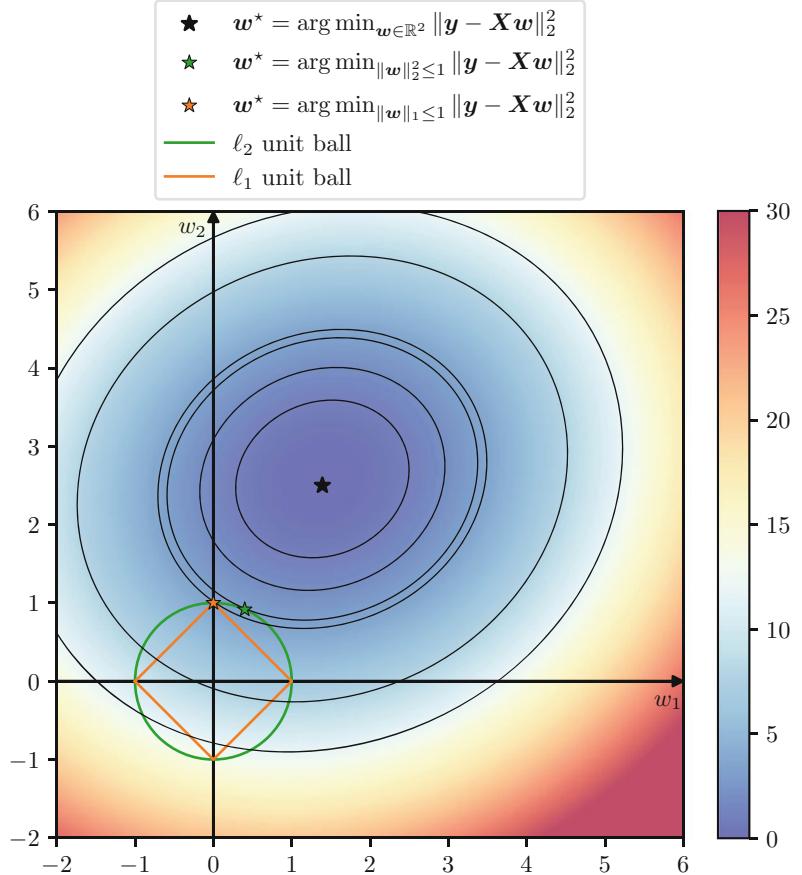


Fig. 7 Illustration of the minimization problem with a constraint on the penalty term. The plot represents the value of the loss function for different values of the two coefficients for a linear regression task. The black star indicates the optimal solution with no constraint. The green and orange stars indicate the optimal solutions when imposing a constraint on the ℓ_2 and ℓ_1 norms of the parameters \mathbf{w} , respectively

mean loss value is dominating the regularization term, making the prediction function fitting too well the training data (overfitting). A good balance between the mean loss value and the regularization term is required to learn the best function.

Since linear regression is one of the oldest and best-known models, the aforementioned penalties were originally introduced for linear regression:

- Linear regression with the ℓ_2 penalty is also known as ridge [7]:

$$\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

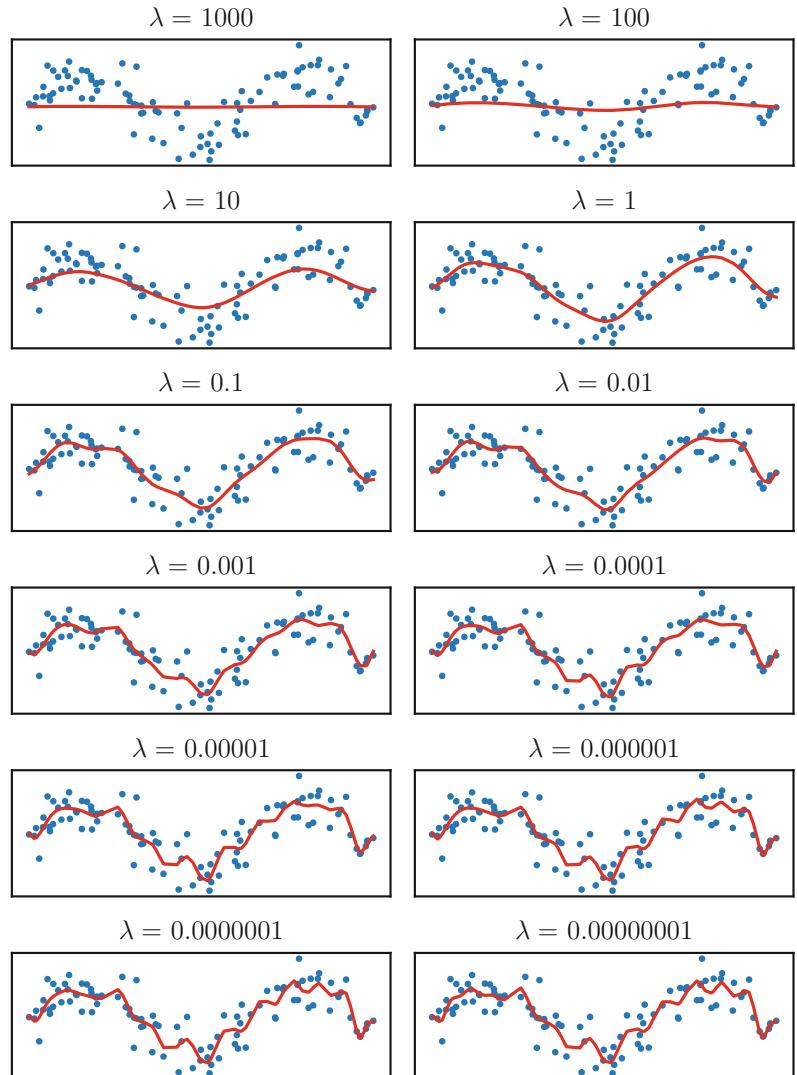


Fig. 8 Illustration of regularization. A kernel ridge regression algorithm is fitted on the training data (blue points) with different values of λ , which is the weight of the regularization in the cost function. The smaller the values of λ , the smaller the weight of the ℓ_2 regularization. The algorithm underfits (respectively, overfits) the data when the value of λ is too large (respectively, low)

As in ordinary least squares regression, there exists a closed formula for the optimal solution:

$$\boldsymbol{w}^* = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

- Linear regression with the ℓ_1 penalty is also known as lasso [8]:

$$\min_{\boldsymbol{w}} \|\mathbf{y} - \mathbf{X}\boldsymbol{w}\|_2^2 + \lambda \|\boldsymbol{w}\|_1$$

- Linear regression with the elastic-net penalty is also known as elastic-net [9]:

$$\min_{\boldsymbol{w}} \|\mathbf{y} - \mathbf{X}\boldsymbol{w}\|_2^2 + \lambda\alpha \|\boldsymbol{w}\|_1 + \lambda(1-\alpha) \|\boldsymbol{w}\|_2^2$$

The penalties can also be added in other models such as logistic regression, support vector machines, artificial neural networks, etc.

8 Support Vector Machine

Linear and logistic regression take into account every training sample in order to find the best line, which is due to their corresponding loss functions: the squared error is zero only if the true and predicted outputs are equal, and the logistic loss is always positive. One could argue that the training samples whose outputs are “easily” well predicted are not relevant: only the training samples whose outputs are not “easily” well predicted or are wrongly predicted should be taken into account. The support vector machine (SVM) is based on this principle (please see Box 4 for an overview of the SVM).

Box 4: Support Vector Machine

- **Main idea:** hyperplane (i.e., line) that maximizes the margin (i.e., the distance between the hyperplane and the closest inputs to the hyperplane).
- **Support vectors:** only the misclassified inputs and the inputs well classified but with low confidence are taken into account.
- **Non-linearity:** decision function can be non-linear with the use of non-linear kernels.
- **Regularization:** ℓ_2 penalty.

8.1 Original Formulation

The original support vector machine was invented in 1963 and was a linear binary classification method [10]. Figure 9 illustrates the main concept of its original version. When both classes are linearly

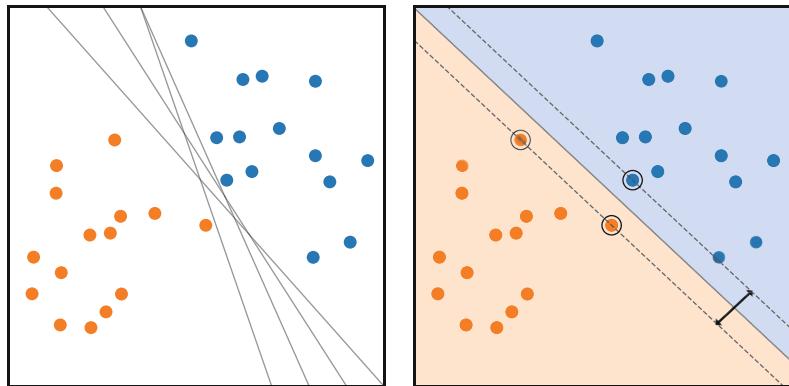


Fig. 9 Support vector machine classifier with linearly separable classes. When two classes are linearly separable, there exist an infinite number of hyperplanes separating them (left). The decision function of the support vector machine classifier is the hyperplane that maximizes the margin, that is, the distance between the hyperplane and the closest points to the hyperplane (right). Support vectors are highlighted with a black circle surrounding them

separable, there exist an infinite number of hyperplanes that separate both classes. The SVM finds the hyperplane that maximizes the *margin*, that is, the distance between the hyperplane and the closest points of both classes to the hyperplane, while linearly separating both classes.

The SVM was later updated to non-separable classes [11]. Figure 10 illustrates the role of the margin in this case. The dashed lines correspond to the hyperplanes defined by the equations $\mathbf{x}^\top \mathbf{w} = +1$ and $\mathbf{x}^\top \mathbf{w} = -1$. The margin is the distance between both hyperplanes and is equal to $2/\|\mathbf{w}\|_2^2$. It defines which samples are included in the decision function of the model: a sample is included if and only if it is inside the margin or outside the margin and misclassified. Such samples are called *support vectors* and are illustrated in Fig. 10 with a black circle surrounding them. In this case, the margin can be seen a regularization term: the larger the margin is, the more support vectors are included in the decision function, the more regularized the model is.

The loss function for the SVM is called the *hinge loss* and is defined as:

$$\ell_{\text{hinge}}(y, f(\mathbf{x})) = \max(0, 1 - yf(\mathbf{x}))$$

Figure 11 illustrates the curves of the logistic and hinge losses. The logistic loss is always positive, even when the point is accurately classified with high confidence (i.e., when $yf(\mathbf{x}) \gg 0$), whereas the hinge loss is equal to zero when the point is accurately classified with good confidence (i.e., when $yf(\mathbf{x}) \geq 1$). One can see that a sample (\mathbf{x}, y) is a support vector if and only if $yf(\mathbf{x}) \geq 1$, that is, if and only if $\ell_{\text{hinge}}(y, f(\mathbf{x})) = 0$.

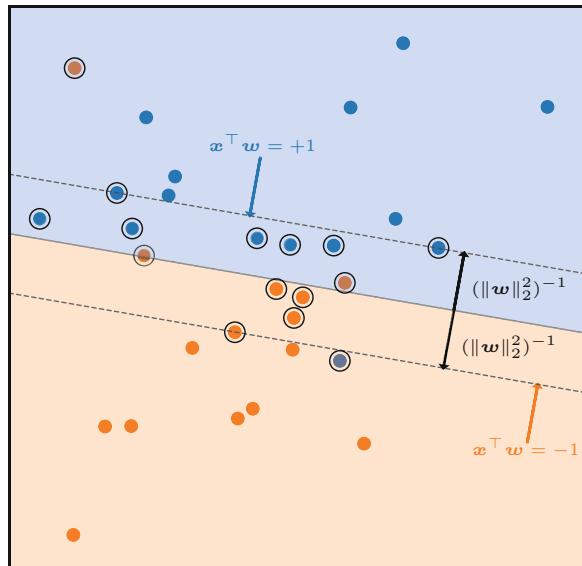


Fig. 10 Decision function of a support vector machine classifier with a linear kernel when both classes are not strictly linearly separable. The support vectors are the training points within the margin of the decision function and the misclassified training points. The support vectors are highlighted with a black circle surrounding them

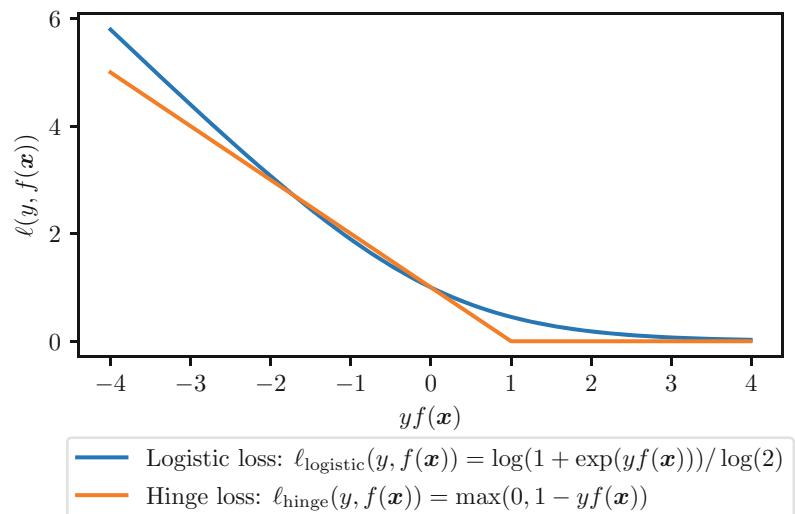


Fig. 11 Binary classification losses. The logistic loss is always positive, even when the point is accurately classified with high confidence (i.e., when $yf(\mathbf{x}) \gg 0$), whereas the hinge loss is equal to zero when the point is accurately classified with good confidence (i.e., when $yf(\mathbf{x}) \geq 1$)

The optimal \mathbf{w} coefficients for the original version are estimated by minimizing an objective function consisting of the sum of the *hinge loss* values and a ℓ_2 penalty term (which is inversely proportional to the margin):

$$\min_{\mathbf{w}} \sum_{i=1}^n \max(0, 1 - y^{(i)} \mathbf{x}^{(i)\top} \mathbf{w}) + \frac{1}{2C} \|\mathbf{w}\|_2^2$$

8.2 General Formulation with Kernels

The SVM was later updated to non-linear decision functions with the use of *kernels* [12].

In order to have a non-linear decision function, one could map the input space I into another space (often called the *feature space*), denoted by \mathcal{G} , using a function denoted by ϕ :

$$\begin{aligned} \phi : I &\rightarrow \mathcal{G} \\ \mathbf{x} &\mapsto \phi(\mathbf{x}) \end{aligned}$$

The decision function would still be linear (with a dot product), but in the feature space:

$$f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w}$$

Unfortunately, solving the corresponding minimization problem is not trivial:

$$\min_{\mathbf{w}} \sum_{i=1}^n \max(0, 1 - y^{(i)} \phi(\mathbf{x}^{(i)})^\top \mathbf{w}) + \frac{1}{2C} \|\mathbf{w}\|_2^2 \quad (3)$$

Nonetheless, two mathematical properties make the use of non-linear transformations in the feature space possible: the *kernel trick* and the *representer theorem*.

The kernel trick asserts that the dot product in the feature space can be computed using only the points from the input space and a *kernel function*, denoted by K :

$$\forall \mathbf{x}, \mathbf{x}' \in I, \phi(\mathbf{x})^\top \phi(\mathbf{x}') = K(\mathbf{x}, \mathbf{x}')$$

The representer theorem [13, 14] asserts that, under certain conditions on the kernel K and the feature space \mathcal{G} associated with the function ϕ , any minimizer of Eq. 3 admits the following form:

$$f = \sum_{i=1}^n \alpha_i K(\cdot, \mathbf{x}^{(i)})$$

where $\boldsymbol{\alpha}$ solves:

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^n \max(0, 1 - y^{(i)} [\mathbf{K}\boldsymbol{\alpha}]_i) + \frac{1}{2C} \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$$

where \mathbf{K} is the $n \times n$ matrix consisting of the evaluations of the kernel on all the pairs of training samples: $\forall i, j \in \{1, \dots, n\}$, $K_{ij} = K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$.

Because the hinge loss is equal to zero if and only if $yf(\mathbf{x})$ is greater than or equal to 1, only the training samples $(\mathbf{x}^{(i)}, y^{(i)})$ such that $y^{(i)}f(\mathbf{x}^{(i)}) < 1$ have a nonzero α_i coefficient. These points are the so-called support vectors, and this is why they are the only training samples contributing to the decision function of the model:

$$\text{SV} = \{i \in \{1, \dots, n\} \mid \alpha_i \neq 0\}$$

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)}) = \sum_{i \in \text{SV}} \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$$

The kernel trick and the representer theorem show that it is more practical to work with the kernel K instead of the mapping function ϕ . Popular kernel functions include:

- The linear kernel:

$$K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$$

- The polynomial kernel:

$$K(\mathbf{x}, \mathbf{x}') = (\gamma \mathbf{x}^\top \mathbf{x}' + c_0)^d \quad \text{with } \gamma > 0, c_0 \geq 0, d \in \mathbb{N}^*$$

- The sigmoid kernel:

$$K(\mathbf{x}, \mathbf{x}') = \tanh(\gamma \mathbf{x}^\top \mathbf{x}' + c_0) \quad \text{with } \gamma > 0, c_0 \geq 0$$

- The radial basis function (RBF) kernel:

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|_2^2) \quad \text{with } \gamma > 0$$

The linear kernel yields a linear decision function and is actually identical to the original formulation of the SVM (one can show that there is a mapping between the α and w coefficients). Non-linear kernels allow for non-linear, more complex, decision functions. This is particularly useful when the data is not linearly separable, which is the most common use case. Figure 12 illustrates the decision function and the margin of a SVM classification model for four different kernels.

The SVM was also extended to regression tasks with the use of the ϵ -insensitive loss. Similar to the hinge loss, which is equal to zero for points that are correctly classified and outside the margin, the ϵ -insensitive loss is equal to zero when the error between the true target value and the predicted value is not greater than ϵ :

$$\ell_{\epsilon-\text{insensitive}}(y, f(\mathbf{x})) = \max(0, |y - f(\mathbf{x})| - \epsilon)$$

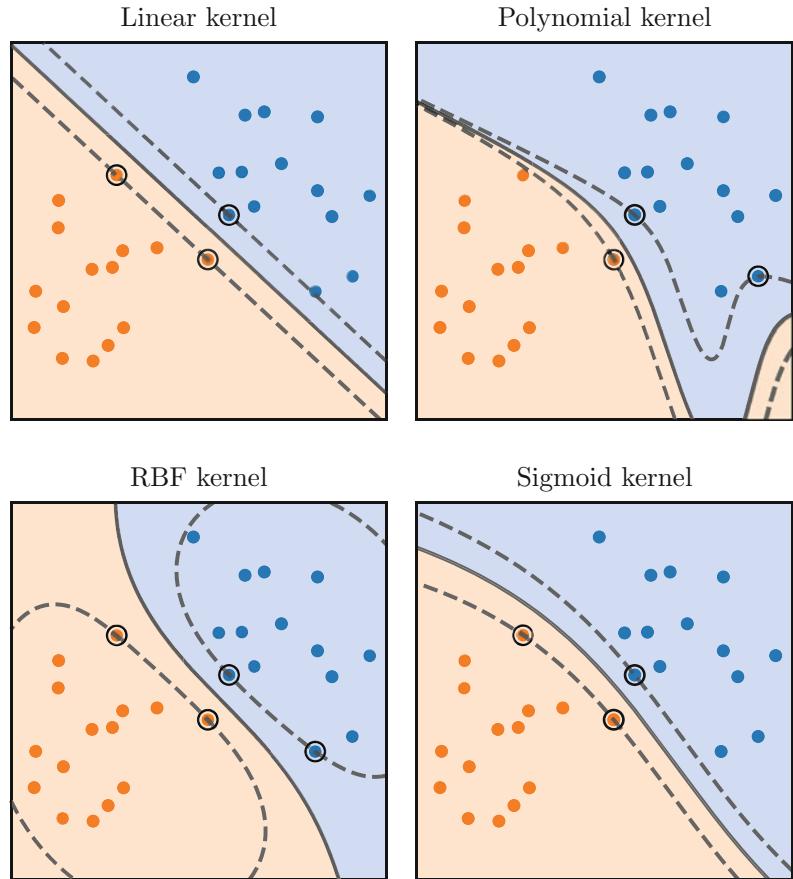


Fig. 12 Impact of the kernel on the decision function of a support vector machine classifier. A non-linear kernel allows for a non-linear decision function

The objective function for the SVM regression method combines the values of ε -insensitive loss of the training points and the ℓ_2 penalty:

$$\min_{\mathbf{w}} \sum_{i=1}^n \max \left(0, |y^{(i)} - \phi(\mathbf{x}^{(i)})^\top \mathbf{w}| - \varepsilon \right) + \frac{1}{2C} \|\mathbf{w}\|_2^2$$

Figure 13 illustrates the curves of three regression losses. The squared error loss takes very small values for small errors and very high values for high errors, whereas the absolute error loss takes small values for small errors and high values for high errors. Both losses take small but nonzero values when the error is small. On the contrary, the ε -insensitive loss is null when the error is small and otherwise equal to the absolute error loss minus ε .

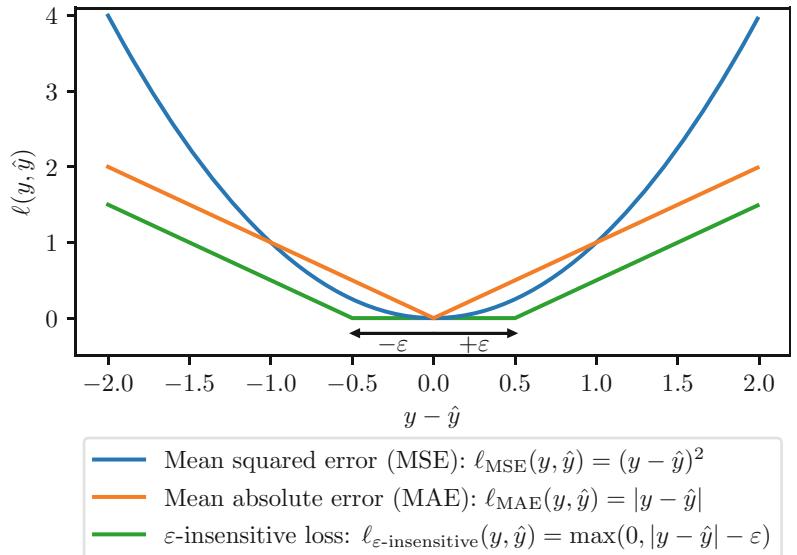


Fig. 13 Regression losses. The squared error loss takes very small values for small errors and very large values for large errors, whereas the absolute error loss takes small values for small errors and large values for large errors. Both losses take small but nonzero values when the error is small. On the contrary, the ε -insensitive loss is null when the error is small and otherwise equal the absolute error loss minus ε . When computed over several samples, the squared and absolute error losses are often referred to as mean squared error (MSE) and mean absolute error (MAE), respectively

9 Multiclass Classification

The classification methods that we presented so far, logistic regression and support vector machines, are binary classifiers: they can only be used when there are only two possible outcomes. However, in practice, it is common to have more than two possible outcomes. For instance, differential diagnosis of brain disorders is often between several, and not only two, diseases.

Several strategies have been proposed to extend any binary classification method to multiclass classification tasks. They all rely on transforming the multiclass classification task into several binary classification tasks. In this section, we present the most commonly used strategies: *one-vs-rest*, *one-vs-one*, and *error correcting output code* [15]. Figure 14 illustrates the main ideas of these approaches. But first, we present a natural extension of logistic regression to multiclass classification tasks which is often referred to as *multinomial logistic regression* [5].

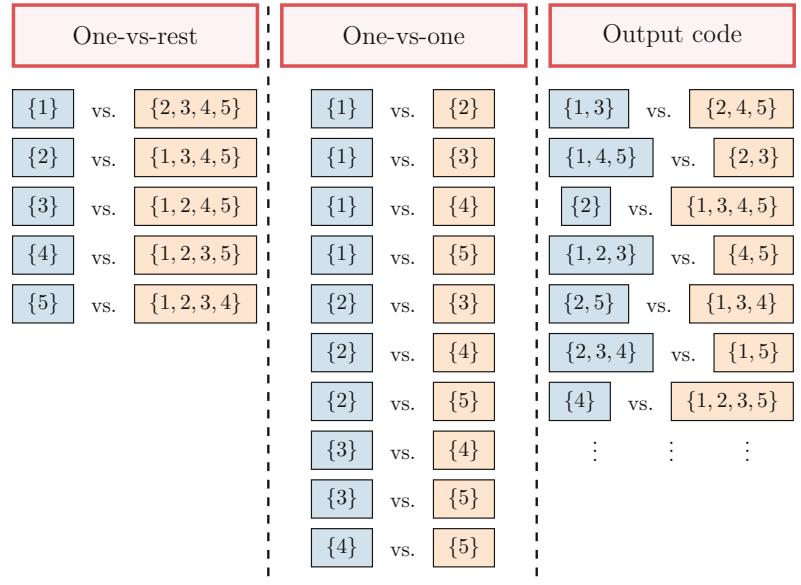


Fig. 14 Main approaches to convert a multiclass classification task into several binary classification tasks. In the one-vs-rest approach, each class is associated with a binary classification model that is trained to separate this class from all the other classes. In the one-vs-one approach, a binary classifier is trained on each pair of classes. In the error correcting output code approach, the classes are (randomly) split into two groups, and a binary classifier is trained for each split

9.1 Multinomial Logistic Regression

For binary classification, logistic regression is characterized by a hyperplane: the signed distance to the hyperplane is mapped into the probability of belonging to the positive class using the sigmoid function. However, for multiclass classification, a single hyperplane is not enough to characterize all the classes. Instead, each class C_k is characterized by a hyperplane \mathbf{w}_k , and, for any input \mathbf{x} , one can compute the signed distance $\mathbf{x}^\top \mathbf{w}_k$ between the input \mathbf{x} and the hyperplane \mathbf{w}_k . The signed distances are mapped into probabilities using the softmax function, defined as

$$\text{softmax}(x_1, \dots, x_q) = \left(\frac{\exp(x_1)}{\sum_{j=1}^q \exp(x_j)}, \dots, \frac{\exp(x_q)}{\sum_{j=1}^q \exp(x_j)} \right), \text{ as follows:}$$

$$\forall k \in \{1, \dots, q\}, P(y = C_k | \mathbf{x} = \mathbf{x}) = \frac{\exp(\mathbf{x}^\top \mathbf{w}_k)}{\sum_{j=1}^q \exp(\mathbf{x}^\top \mathbf{w}_j)}$$

The coefficients $(\mathbf{w}_k)_{1 \leq k \leq q}$ are still estimated by maximizing the likelihood function:

$$L(\mathbf{w}_1, \dots, \mathbf{w}_q) = \prod_{i=1}^n \prod_{k=1}^q P(y = C_k | \mathbf{x} = \mathbf{x}^{(i)})^{\mathbf{1}_{y^{(i)} = c_k}}$$

which is equivalent to minimizing the negative log-likelihood:

$$\begin{aligned}
& - \log(L(\mathbf{w}_1, \dots, \mathbf{w}_q)) \\
&= - \sum_{i=1}^n \sum_{k=1}^q \mathbf{1}_{y^{(i)} = C_k} \log(P(y = C_k | \mathbf{x} = \mathbf{x}^{(i)})) \\
&= \sum_{i=1}^n - \sum_{k=1}^q \mathbf{1}_{y^{(i)} = C_k} \log \left(\frac{\exp(\mathbf{x}^{(i)\top} \mathbf{w}_k)}{\sum_{j=1}^q \exp(\mathbf{x}^{(i)\top} \mathbf{w}_j)} \right) \\
&= \sum_{i=1}^n \ell_{\text{cross-entropy}}(y^{(i)}, \text{softmax}(\mathbf{x}^{(i)\top} \mathbf{w}_1, \dots, \mathbf{x}^{(i)\top} \mathbf{w}_q))
\end{aligned}$$

where $\ell_{\text{cross-entropy}}$ is known as the *cross-entropy loss* and is defined, for any label y and any vector of probabilities (π_1, \dots, π_q) , as:

$$\ell_{\text{cross-entropy}}(y, (\pi_1, \dots, \pi_q)) = - \sum_{k=1}^q \mathbf{1}_{y = C_k} \log \pi_k$$

This loss is commonly used to train artificial neural networks on classification tasks and is equivalent to the logistic loss in the binary case.

Figure 15 illustrates the impact of the strategy used to handle a multiclass classification task on the decision function.

9.2 One-vs-Rest

A strategy to transform a multiclass classification task into several binary classification tasks is to fit a binary classifier for each class: the positive class is the given class, and the negative class consists of all the other classes merged into a single class. This strategy is known as *one-vs-rest*. The advantage of this strategy is that each class is characterized by a single model, so that it is possible to gain deeper knowledge about the class by inspecting its corresponding model. A consequence is that the predictions for new samples take into account the confidence of the models: the predicted class for a new input is the class for which the corresponding model is the most confident that this input belongs to its class. The one-vs-rest strategy is the most commonly used strategy and usually a good default choice.

9.3 One-vs-One

Another strategy is to fit a binary classifier for each pair of classes: this strategy is known as *one-vs-one*. The advantage of this strategy is that the classes in each binary classification task are “pure”, in the sense that different classes are never merged into a single class. However, the number of binary classifiers that needs to be trained is larger for the one-vs-one strategy ($\frac{1}{2}q(q-1)$) than for the one-vs-rest strategy (q). Nonetheless, for the one-vs-one strategy, the number of training samples in each binary classification task is smaller than the total number of samples, which makes training each binary classifier usually faster. Another drawback is that this strategy is less interpretable compared to the one-vs-rest strategy, as the predicted class corresponds to the class obtaining the most

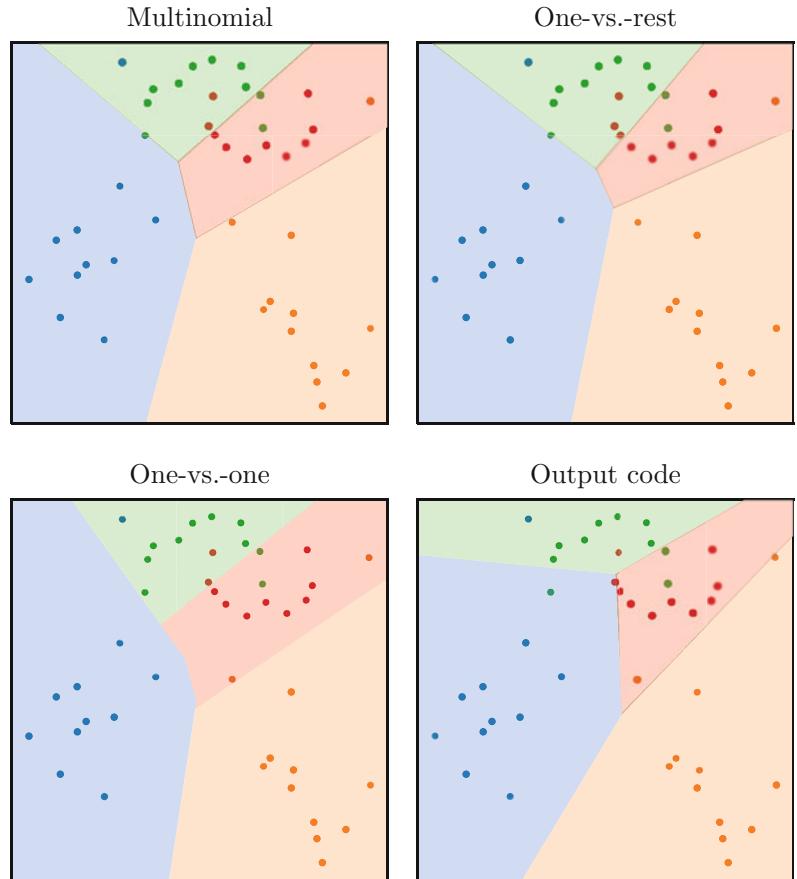


Fig. 15 Illustration of the impact of the strategy used to handle a multiclass classification task on the decision function of a logistic regression model

votes (i.e., winning the most one-vs-one matchups), which does not take into account the confidence in winning each matchup.¹ For instance, winning a one-vs-one matchup with 0.99 probability gives the same result as winning the same matchup with 0.51 probability, i.e., one vote.

9.4 Error Correcting Output Code

A substantially different strategy, inspired by the theory of error correction code, consists in merging a subset of classes into one class and the other subset into the other class, for each binary classification task. This data is often called the code book and can be represented as a matrix whose rows correspond to the classes and whose columns correspond to the binary classification tasks. The matrix consists only of -1 and $+1$ values that represent the corresponding label for each class and for each binary task.² For

¹ The confidences are actually taken into account but only in the event of a tie.

² The values are 0 and 1 when the classifier does not return scores but only probabilities.

any input, each binary classifier returns the score (or probability) associated with the positive class. The predicted class for this input is the class whose corresponding vector is the most similar to the vector of scores, with similarity being assessed with the Euclidean distance (the lower, the more similar). There exist advanced strategies to define the code book, but it has been argued than a random code book usually gives as good results as a sophisticated one [16].

10 Decision Functions with Normal Distributions

Normal distributions are popular distributions because they are commonly found in nature. For instance, the distribution of heights and birth weights of human beings can be approximated using normal distributions. Moreover, normal distributions are particularly easy to work with from a mathematical point of view. For these reasons, a common model consists in assuming that the training input vectors are independently sampled from normal distributions.

A possible classification model consists in assuming that, for each class, all the corresponding inputs are sampled from a normal distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$:

$$\forall i \text{ such that } y^{(i)} = C_k, \quad \mathbf{x}^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Using the probability density function of a normal distribution, one can compute the probability density of any input \mathbf{x} associated with the distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ of class C_k :

$$p_{\mathbf{x}|y=C_k}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2} [\mathbf{x} - \boldsymbol{\mu}_k]^\top \boldsymbol{\Sigma}_k^{-1} [\mathbf{x} - \boldsymbol{\mu}_k]\right)$$

With such a probabilistic model, it is easy to compute the probability that a sample belongs to class C_k using Bayes rule:

$$P(y = C_k | \mathbf{x} = \mathbf{x}) = \frac{p_{\mathbf{x}|y=C_k}(\mathbf{x}) P(y = C_k)}{p_{\mathbf{x}}(\mathbf{x})}$$

With normal distributions, it is mathematically easier to work with log-probabilities:

$$\begin{aligned}
& \log P(y = C_k | \mathbf{x} = \mathbf{x}) \\
&= \log p_{x|y=C_k}(\mathbf{x}) + \log P(y = C_k) - \log p_x(\mathbf{x}) \\
&= -\frac{1}{2} [\mathbf{x} - \boldsymbol{\mu}_k]^\top \boldsymbol{\Sigma}_k^{-1} [\mathbf{x} - \boldsymbol{\mu}_k] - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| + \log P(y = C_k) \\
&\quad - \frac{p}{2} \log(2\pi) - \log p_x(\mathbf{x}) \\
&= -\frac{1}{2} \mathbf{x}^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x} + \mathbf{x}^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \\
&\quad - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| + \log P(y = C_k) \\
&\quad - \frac{p}{2} \log(2\pi) - \log p_x(\mathbf{x})
\end{aligned} \tag{4}$$

It is also possible to make further assumptions on the covariance matrices that lead to different models. In this section, we present the most commonly used ones: naive Bayes, linear discriminant analysis, and quadratic discriminant analysis. Figure 16 illustrates the covariance matrices and the decision functions for these models in the two-dimensional case.

10.1 Naive Bayes

The naive Bayes model assumes that, conditionally to each class C_k , the features are independent and have the same variance σ_k^2 :

$$\forall k, \boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}_p$$

Equation 4 can thus be further simplified:

$$\begin{aligned}
& \log P(y = C_k | \mathbf{x} = \mathbf{x}) \\
&= -\frac{1}{2\sigma_k^2} \mathbf{x}^\top \mathbf{x} + \frac{1}{\sigma_k^2} \mathbf{x}^\top \boldsymbol{\mu}_k - \frac{1}{2\sigma_k^2} \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_k - \log \sigma_k + \log P(y = C_k) \\
&\quad - \frac{p}{2} \log(2\pi) - \log p_x(\mathbf{x}) \\
&= \mathbf{x}^\top \mathbf{W}_k \mathbf{x} + \mathbf{x}^\top \mathbf{w}_k + w_{0k} + s
\end{aligned}$$

where:

- $\mathbf{W}_k = -\frac{1}{2\sigma_k^2} \mathbf{I}_p$ is the matrix of the quadratic term for class C_k .
- $\mathbf{w}_k = \frac{1}{\sigma_k^2} \boldsymbol{\mu}_k$ is the vector of the linear term for class C_k .
- $w_{0k} = -\frac{1}{2\sigma_k^2} \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_k - \log \sigma_k + \log P(y = C_k)$ is the intercept for class C_k .
- $s = -\frac{p}{2} \log(2\pi) - \log p_x(\mathbf{x})$ is a term that does not depend on class C_k .

Therefore, naive Bayes is a quadratic model. The probabilities for input \mathbf{x} to belong to each class C_k can then easily be computed:

$$P(y = C_k | \mathbf{x} = \mathbf{x}) = \frac{\exp(\mathbf{x}^\top \mathbf{W}_k \mathbf{x} + \mathbf{x}^\top \mathbf{w}_k + w_{0k})}{\sum_{j=1}^k \exp(\mathbf{x}^\top \mathbf{W}_j \mathbf{x} + \mathbf{x}^\top \mathbf{w}_j + w_{0j})}$$

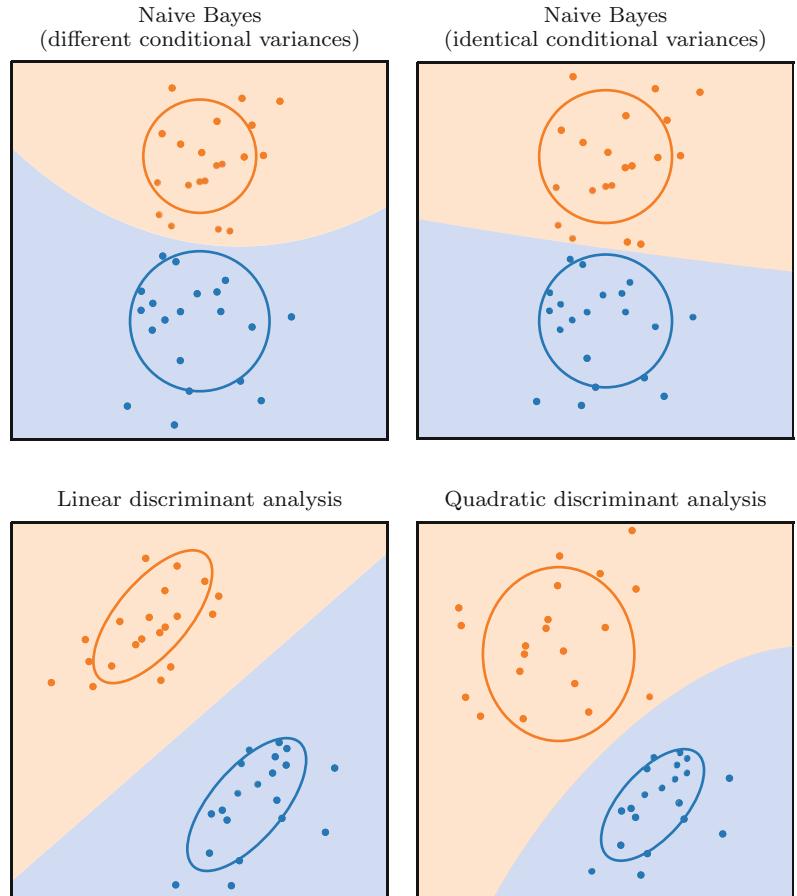


Fig. 16 Illustration of decision functions with normal distributions. A two-dimensional covariance matrix can be represented as an ellipse. In the naive Bayes model, the features are assumed to be independent and to have the same variance conditionally to the class, leading to covariance matrices being represented as circles. When the covariance matrices are assumed to be identical, the decision functions are linear instead of quadratic

With the naive Bayes model, it is relatively common to have the conditional variances σ_k^2 to all be equal:

$$\forall k, \Sigma_k = \sigma_k^2 I_p = \sigma^2 I_p$$

In this case, Eq. 4 can be even further simplified:

$$\begin{aligned}
 & \log P(y = C_k | \mathbf{x} = \mathbf{x}) \\
 &= -\frac{1}{2\sigma^2} \mathbf{x}^\top \mathbf{x} + \frac{1}{\sigma^2} \mathbf{x}^\top \boldsymbol{\mu}_k - \frac{1}{2\sigma^2} \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_k - \log \sigma_k + \log P(y = C_k) \\
 &\quad - \frac{p}{2} \log(2\pi) - \log p_x(\mathbf{x}) \\
 &= \mathbf{x}^\top \mathbf{w}_k + w_{0k} + s
 \end{aligned}$$

where:

- $\mathbf{w}_k = \frac{1}{\sigma^2} \boldsymbol{\mu}_k$ is the vector of the linear term for class C_k .
- $w_{0k} = -\frac{1}{2\sigma^2} \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_k + \log P(y = C_k)$ is the intercept for class C_k .
- $s = -\frac{1}{2\sigma^2} \mathbf{x}^\top \mathbf{x} - \log \sigma - \frac{p}{2} \log(2\pi) - \log p_x(\mathbf{x})$ is a term that does not depend on class C_k .

In this case, naive Bayes becomes a linear model.

10.2 Linear Discriminant Analysis

Linear discriminant analysis (LDA) makes the assumption that all the covariance matrices are identical but otherwise arbitrary:

$$\forall k, \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$$

Therefore, Eq. 4 can be further simplified:

$$\begin{aligned} & \log P(y = C_k | \mathbf{x} = \mathbf{x}) \\ &= -\frac{1}{2} [\mathbf{x} - \boldsymbol{\mu}_k]^\top \boldsymbol{\Sigma}^{-1} [\mathbf{x} - \boldsymbol{\mu}_k] - \frac{1}{2} \log |\boldsymbol{\Sigma}| + \log P(y = C_k) \\ &\quad - \frac{p}{2} \log(2\pi) - \log p_x(\mathbf{x}) \\ &= -\frac{1}{2} (\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} - \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} + \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k) \\ &\quad - \frac{1}{2} \log |\boldsymbol{\Sigma}| + \log P(y = C_k) - \frac{p}{2} \log(2\pi) - \log p_x(\mathbf{x}) \\ &= -\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log P(y = C_k) - \frac{1}{2} \log |\boldsymbol{\Sigma}| \\ &\quad - \frac{p}{2} \log(2\pi) - \log p_x(\mathbf{x}) \\ &= \mathbf{x}^\top \mathbf{w}_k + w_{0k} + s \end{aligned}$$

where:

- $\mathbf{w}_k = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k$ is the vector of coefficients for class C_k .
- $w_{0k} = -\frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log P(y = C_k)$ is the intercept for class C_k .
- $s = -\frac{1}{2} \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{p}{2} \log(2\pi) - \log p_x(\mathbf{x})$ is a term that does not depend on class C_k .

Therefore, linear discriminant analysis is a linear model. When $\boldsymbol{\Sigma}$ is diagonal, linear discriminant analysis is identical to naive Bayes with identical conditional variances.

The probabilities for input \mathbf{x} to belong to each class C_k can then easily be computed:

$$P(y = C_k | \mathbf{x} = \mathbf{x}) = \frac{\exp(\mathbf{x}^\top \mathbf{w}_k + w_{0k})}{\sum_{j=1}^k \exp(\mathbf{x}^\top \mathbf{w}_j + w_{0j})}$$

10.3 Quadratic Discriminant Analysis

Quadratic discriminant analysis makes no assumption on the covariance matrices $\boldsymbol{\Sigma}_k$ that can all be arbitrary. Equation 4 can be written as:

$$\begin{aligned}
& \log P(y = C_k | \mathbf{x} = \mathbf{x}) \\
&= -\frac{1}{2} \mathbf{x}^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x} + \mathbf{x}^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| \\
&\quad + \log P(y = C_k) - \frac{p}{2} \log(2\pi) - \log p_{\mathbf{x}}(\mathbf{x}) \\
&= \mathbf{x}^\top \mathbf{W}_k \mathbf{x} + \mathbf{x}^\top \mathbf{w}_k + w_{0k} + s
\end{aligned}$$

where:

- $\mathbf{W}_k = -\frac{1}{2} \boldsymbol{\Sigma}_k^{-1}$ is the matrix of the quadratic term for class C_k .
- $\mathbf{w}_k = \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k$ is the vector of the linear term for class C_k .
- $w_{0k} = -\frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| + \log P(y = C_k)$ is the intercept for class C_k .
- $s = -\frac{p}{2} \log(2\pi) - \log p_{\mathbf{x}}(\mathbf{x})$ is a term that does not depend on class C_k .

Therefore, quadratic discriminant analysis is a quadratic model.

The probabilities for input \mathbf{x} to belong to each class C_k can then easily be computed:

$$P(y = C_k | \mathbf{x} = \mathbf{x}) = \frac{\exp(\mathbf{x}^\top \mathbf{W}_k \mathbf{x} + \mathbf{x}^\top \mathbf{w}_k + w_{0k})}{\sum_{j=1}^k \exp(\mathbf{x}^\top \mathbf{W}_j \mathbf{x} + \mathbf{x}^\top \mathbf{w}_j + w_{0j})}$$

11 Tree-Based Methods

11.1 Decision Tree

Binary decisions based on conditional statements are frequently used in everyday life because they are intuitive and easy to understand. Figure 17 illustrates a general approach when someone is ill. Depending on conditional statements (severity of symptoms, ability to quickly consult a specialist), the decision (consult your general practitioner or a specialist, or call for emergency services) is different. Models with such an architecture are often used in machine learning and are called *decision trees*.

A decision tree is an algorithm containing only conditional statements and can be represented with a tree [17]. This graph consists of:

- Decision nodes for all the conditional statements
- Branches for the potential outcomes of each decision node
- Leaf nodes for the final decision

Figure 18 illustrates a decision tree and its corresponding decision function. For a given sample, the final decision is obtained by following its corresponding path, starting at the root node.

A decision tree recursively partitions the feature space in order to group samples with the same labels or similar target values. At each node, the objective is to find the best (feature, threshold) pair so that both subsets obtained with this split are the most *pure*, that

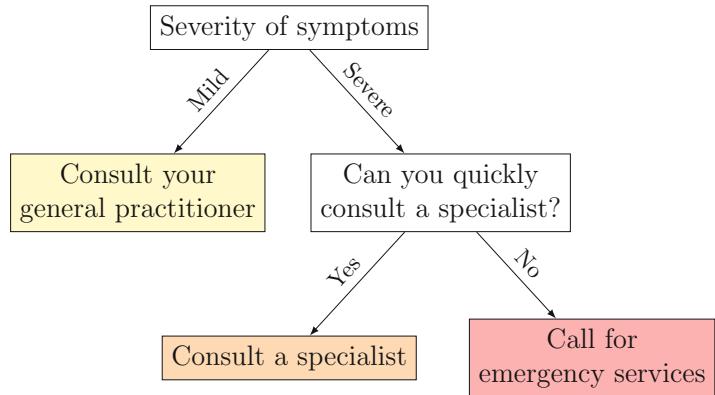


Fig. 17 A general thought process when being ill. Depending on conditional statements (severity of symptoms, ability to quickly consult a specialist), the decision (consult your general practitioner or a specialist, or call for emergency services) is different

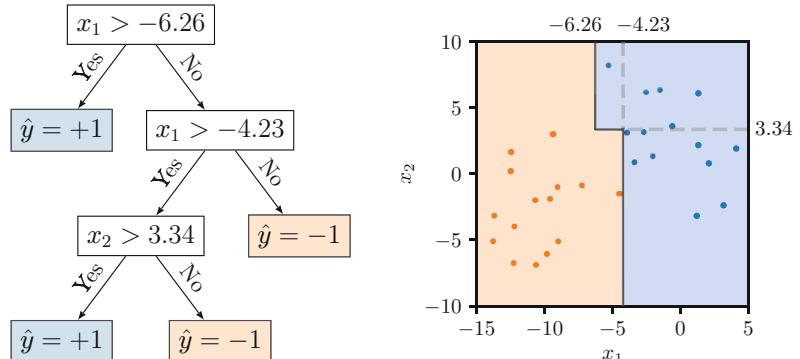


Fig. 18 A decision tree: (left) the rules learned by the decision tree and (right) the corresponding decision function

is, homogeneous. To do so, the best (feature, threshold) pair is defined as the pair that minimizes an *impurity* criterion.

Let $\mathcal{S} \subseteq \mathcal{X}$ be a subset of training samples. For classification tasks, the distribution of the classes, that is, the proportion of each class, is used to measure the purity of the subset. Let p_k be the proportion of samples from class C_k in a given partition:

$$p_k = \frac{1}{|\mathcal{S}|} \sum_{y \in \mathcal{S}} \mathbf{1}_{y=C_k}$$

Popular impurity criteria for classification tasks include:

- Gini index: $\sum k p_k (1 - p_k)$
- Entropy: $-\sum_k p_k \log(p_k)$
- Misclassification: $1 - \max_k p_k$

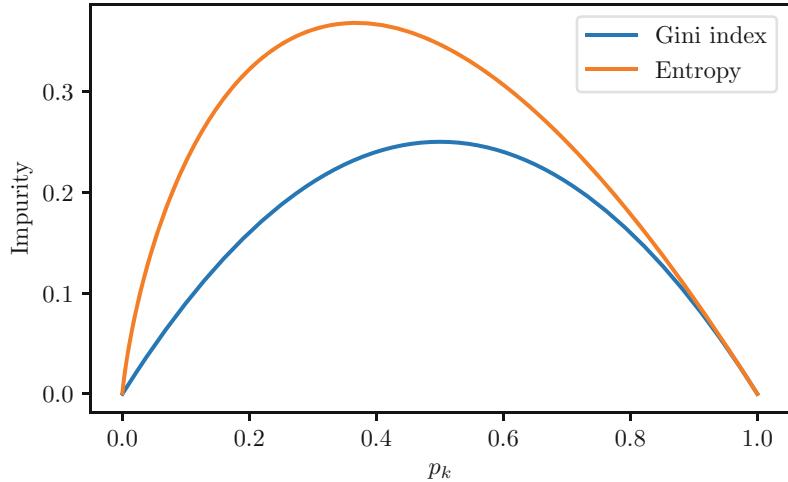


Fig. 19 Illustration of Gini index and entropy. The entropy function takes larger values than the Gini index, especially for $p_k < 0.8$, which thus is more discriminative against heterogeneous subsets (when most classes only represent only a small proportion of the samples) than Gini index

Figure 19 illustrates the values of the Gini index and the entropy for a single class C_k and for different proportions of samples p_k . One can see that the entropy function takes larger values than the Gini index, especially for $p_k < 0.8$. Since the sum of the proportions is equal to 1, most classes only represent a small proportion of the samples. Therefore, a simple interpretation is that entropy is more discriminative against heterogeneous subsets than the Gini index. Misclassification only takes into account the proportion of the most common class and tends to be even less discriminative against heterogeneous subsets than both entropy and Gini index.

For regression tasks, the mean error from a reference value (such as the mean or the median) is often used as the impurity criterion:

- Mean squared error: $\frac{1}{|S|} \sum_{y \in S} (y - \bar{y})^2$ with $\bar{y} = \frac{1}{|S|} \sum_{y \in S} y$
- Mean absolute error: $\frac{1}{|S|} \sum_{y \in S} |y - \text{median}_S(y)|$

Theoretically, a tree can grow until every leaf node is perfectly pure. However, such a tree would have a lot of branches and would be very complex, making it prone to overfitting. Several strategies are commonly used to limit the size of the tree. One approach consists in growing the tree with no restriction and then *pruning* the tree, that is, replacing subtrees with nodes [17]. Other popular strategies to limit the complexity of the tree are usually applied while the tree is grown and include setting:

- A maximum depth for the tree
- A minimum number of samples required to be at an internal node

- A minimum number of samples required to split a given partition
- A maximum number of leaf nodes
- A maximum number of features considered (instead of all the features) to find the best split
- A minimum impurity decrease to split an internal node

11.2 Random Forest

One limitation of decision trees is their simplicity. Decision trees tend to use a small fraction of the features in their decision function. In order to use more features in the decision tree, growing a larger tree is required, but large trees tend to suffer from overfitting, that is, having a low bias but a high variance. One solution to decrease the variance without much increasing the bias is to build an ensemble of trees with randomness, hence the name *random forest* [18]. An overview of random forests can be found in Box 5.

In a bid to have trees that are not perfectly correlated (thus building actually different trees), each tree is built using only a subset of the training samples obtained with random sampling. Moreover, for each decision node of each tree, only a subset of the features are considered to find the best split.

The final prediction is obtained by averaging the predictions of each tree. For classification tasks, the predicted class is either the most commonly predicted class (hard-voting) or the one with the highest mean probability estimate (soft-voting) across the trees. For regression tasks, the predicted value is usually the mean of the predicted values across the trees.

Box 5: Random Forest

- **Random forest:** ensemble of decision trees with randomness introduced to build different trees
- **Decision tree:** algorithm containing only conditional statements and represented with a tree
- **Regularization:** maximum depth for each tree, minimum number of samples required to split a given partition, etc.

11.3 Extremely Randomized Trees

Even though random forests involve randomness in sampling both the samples and the features, trees inside a random forest tend to be correlated, thus limiting the variance decrease. In order to decrease even more the variance of the model (while possibly increasing its bias) by growing less correlated trees, *extremely randomized trees* introduce more randomness [19]. Instead of looking for the best split among all the candidate (feature,

threshold) pairs, one threshold is drawn at random for each candidate feature, and the best of these randomly generated thresholds is chosen as the splitting rule.

12 Clustering

So far, we have presented classic machine learning methods for classification and regression, which are the main components of supervised learning. Each input $x^{(i)}$ had an associated output $y^{(i)}$. In this section, we present clustering, which is an unsupervised machine learning task. In unsupervised learning, only the inputs $x^{(i)}$ are available, with no associated outputs. As the ground truth is not available, the objective is to extract information from the input data without supervising the learning process with the output data.

Clustering consists in finding groups of samples such that:

- Samples from the same group are similar.
- Samples from different groups are different.

For instance, clustering can be used to identify disease subtypes for heterogeneous diseases such as Alzheimer’s disease and Parkinson’s disease.

In this section, we present two of the most common clustering methods: the k -means algorithm and the Gaussian mixture model.

12.1 k -means

The k -means algorithm divides a set of n samples, denoted by X , into a set of k disjoint clusters, each denoted by X_j , such that $X = \{X_1, \dots, X_k\}$.

Figure 20 illustrates the concept of this algorithm. Each cluster X_j is characterized by its *centroid*, denoted by μ_j , that is, the mean of the samples in this cluster:

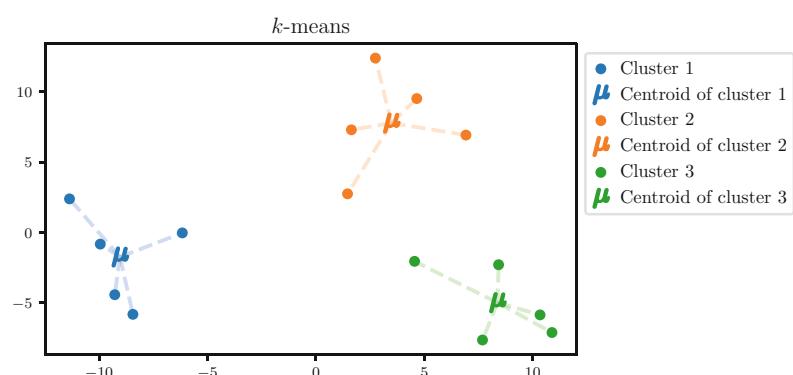


Fig. 20 Illustration of the k -means algorithm. The objective of the algorithm is to find the centroids that minimize the within-cluster sum-of-squares criterion. In this example, the inertia is approximately equal to 184.80 and is the lowest possible inertia, meaning that the represented centroids are optimal

$$\boldsymbol{\mu}_j = \frac{1}{|\mathcal{X}_j|} \sum_{\mathbf{x}^{(i)} \in \mathcal{X}_j} \mathbf{x}^{(i)}$$

The centroids fully define the set of clusters because each sample is assigned to the cluster whose centroid is the closest.

The k -means algorithm aims at finding centroids that minimize the *inertia*, also known as *within-cluster sum-of-squares criterion*:

$$\min_{\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k\}} \sum_{j=1}^k \sum_{\mathbf{x}^{(i)} \in \mathcal{X}_j} \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_j\|_2^2$$

The original algorithm used to find the centroids is often referred to as *Lloyd's algorithm* [20] and is presented in Algorithm 1. After initializing the centroids, a two-step loop is repeated until convergence (when the centroids are identical for two consecutive iterations) consisting of:

1. The *assignment step*, where the clusters are updated based on the current centroids
2. The *update step*, where the centroids are updated based on the current clusters

When clusters are well-defined, a point from a given cluster is likely to stay in this cluster. Therefore, the assignment step can be sped up thanks to the triangle inequality by keeping track of lower and upper bounds for distances between points and centers, at the cost of higher memory usage [21].

Algorithm 1 Lloyd's algorithm (aka naive k -means algorithm)

Result: Centroids $\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k\}$

Initialize the centroids $\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k\}$;

while *not converged* **do**

Assignment step: Compute the clusters (i.e., assign each sample to its nearest centroid):

$$\forall j \in \{1, \dots, k\}, \mathcal{X}_j = \{\mathbf{x}^{(i)} \in \mathcal{X} \mid \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_j\|_2^2 = \min_l \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_l\|_2^2\}$$

Update step: Compute the centroids of the updated clusters:

$$\forall j \in \{1, \dots, k\}, \boldsymbol{\mu}_j = \frac{1}{|\mathcal{X}_j|} \sum_{\mathbf{x}^{(i)} \in \mathcal{X}_j} \mathbf{x}^{(i)}$$

Even though the k -means algorithm is one of the simplest and most used clustering methods, it has several downsides that should be kept in mind.

First, the number of clusters k is a hyperparameter. Setting a value much different from the actual number of clusters may yield poor clusters.

Second, the inertia is not a convex function. Although Lloyd's algorithm is guaranteed to converge, it may converge to a local minimum that is not a global minimum. Figure 21 illustrates the convergence to such centroids. Several strategies are often applied to address this issue, including sophisticated centroid initialization [22] and running the algorithm numerous times and keeping the best run (i.e., the one yielding the lowest inertia).

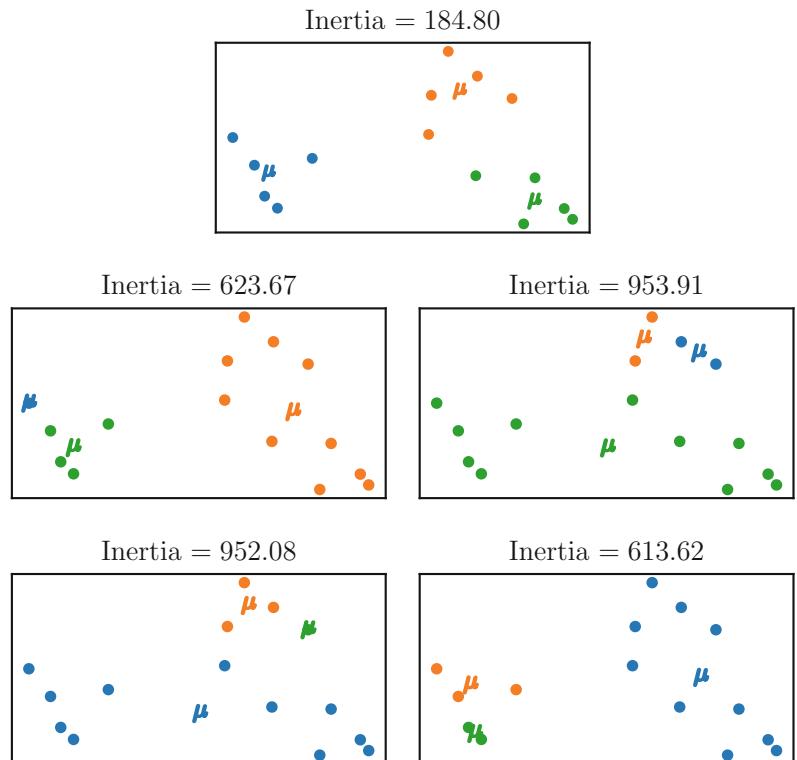


Fig. 21 Illustration of the convergence of the k -means algorithm to bad local minima. In the upper figure, the algorithm converged to a global minimum because the inertia is equal to the minimum possible value (184.80); thus, the obtained clusters are optimal. In the four other figures, the algorithm converged to a local minima that are not global minima because the inertias are higher than the minimum possible value; thus, the obtained clusters are suboptimal

Third, inertia makes the assumption that the clusters are convex and isotropic. The k -means algorithm may yield poor results when this assumption does not hold, such as with elongated clusters or manifolds with irregular shapes.

Fourth, the Euclidean distance tends to be inflated (i.e., the ratio of the distances of the nearest and farthest neighbors to a given target is close to 1) in high-dimensional spaces, making inertia a poor criterion in such spaces [23]. One can alleviate this issue by running a dimensionality reduction method such as principal component analysis prior to the k -means algorithm.

12.2 Gaussian Mixture Model

A mixture model makes the assumption that each sample is generated from a mixture of several independent distributions.

Let k be the number of distributions. Each distribution F_j is characterized by its probability of being picked, denoted by π_j , and its density p_j with parameters θ_j , denoted by $p_j(\cdot; \theta_j)$. Let $\Delta = (\Delta_1, \dots, \Delta_k)$ be a vector-valued random variable such that:

$$\sum_{j=1}^k \Delta_j = 1 \quad \text{and} \quad \forall j \in \{1, \dots, k\}, P(\Delta_j = 1) = 1 - P(\Delta_j = 0) = \pi_j$$

and (x_1, \dots, x_k) be independent random variables such that $x_j \sim F_j$. The samples are assumed to be generated from a random variable x with density p_x such that:

$$x = \sum_{j=1}^k \Delta_j x_j$$

$$\forall x \in \mathcal{X}, p_x(x, \theta) = \sum_{j=1}^k \pi_j p_j(x; \theta_j)$$

A Gaussian mixture model is a particular case of a mixture model in which each distribution F_j is a Gaussian distribution with mean vector μ_j and covariance matrix Σ_j :

$$\forall j \in \{1, \dots, k\}, F_j = \mathcal{N}(\mu_j, \Sigma_j)$$

Figure 22 illustrates the learned distributions from a Gaussian mixture model.

The objective is to find the parameters θ that maximize the likelihood, with $\theta = (\{\mu_j\}_{j=1}^k, \{\Sigma_j\}_{j=1}^k, \{\pi_j\}_{j=1}^k)$:

$$L(\theta) = \prod_{i=1}^n p_X(x^{(i)}; \theta)$$

For computational reasons, it is easier to maximize the log-likelihood:

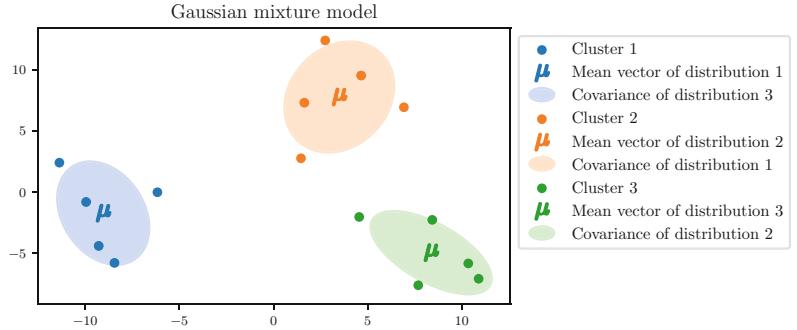


Fig. 22 Gaussian mixture model. For each estimated distribution, the mean vector and the ellipsis consisting of all the points within one standard deviation of the mean are plotted

$$\log(L(\boldsymbol{\theta})) = \sum_{i=1}^n \log(p_X(\mathbf{x}^{(i)}; \boldsymbol{\theta})) = \sum_{i=1}^n \log\left(\sum_{j=1}^k \pi_j p_j(\mathbf{x}; \boldsymbol{\theta}_j)\right)$$

Because the density $p_X(\cdot; \boldsymbol{\theta})$ is a weighted sum of Gaussian densities, the expression cannot be further simplified.

In order to solve this maximization problem, an algorithm called *expectation-maximization* (EM) is often applied [24]. Algorithm 2 describes the main concepts of this algorithm. After initializing the parameters of each distribution, a two-step loop is repeated until convergence (when the parameters are stable over consecutive loops):

- The *expectation step*, in which the probability for each sample $\mathbf{x}^{(i)}$ to have been generated from distribution F_j is computed
- The *maximization step*, in which the probability and the parameters of each distribution are updated

Because it is impossible to know which samples have been generated by each distribution, it is also impossible to directly maximize the log-likelihood, which is why we compute its *expected value* using the posterior probabilities, hence the name *expectation step*. The second step simply consists in maximizing the expected log-likelihood, hence the name *maximization step*.

Algorithm 2 Expectation-maximization algorithm for Gaussian mixture models

Result: Mean vectors $\{\boldsymbol{\mu}_j\}_{j=1}^k$, covariance matrices $\{\boldsymbol{\Sigma}_j\}_{j=1}^k$ and probabilities $\{\pi_j\}_{j=1}^k$

Initialize the mean vectors $\{\boldsymbol{\mu}_j\}_{j=1}^k$, covariance matrices $\{\boldsymbol{\Sigma}_j\}_{j=1}^k$ and probabilities $\{\pi_j\}_{j=1}^k$;

while *not converged* **do**

E-step: Compute the posterior probability $\gamma_i(j)$ for each sample $\mathbf{x}^{(i)}$ to have been generated from distribution F_j :

$$\forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, k\}, \gamma_i(j) = \frac{\pi_j p_j(\mathbf{x}^{(i)}; \boldsymbol{\theta}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^k \pi_l p_l(\mathbf{x}^{(i)}; \boldsymbol{\theta}_l, \boldsymbol{\Sigma}_l)}$$

M-step: Update the parameters of each distribution F_j :

$$\begin{aligned} \forall j \in \{1, \dots, k\}, \boldsymbol{\mu}_j &= \frac{\sum_{i=1}^n \gamma_i(j) \mathbf{x}^{(i)}}{\sum_{i=1}^n \gamma_i(j)} \\ \forall j \in \{1, \dots, k\}, \boldsymbol{\Sigma}_j &= \frac{\sum_{i=1}^n \gamma_i(j) [\mathbf{x}^{(i)} - \boldsymbol{\mu}_j][\mathbf{x}^{(i)} - \boldsymbol{\mu}_j]^\top}{\sum_{i=1}^n \gamma_i(j)} \\ \forall j \in \{1, \dots, k\}, \pi_j &= \frac{1}{n} \sum_{i=1}^n \gamma_i(j) \end{aligned}$$

Lloyd's and EM algorithms have a lot of similarities. In the first step, the assignment step assigns each sample to its closest cluster, whereas the expectation step computes the probability for each sample to have been generated from each distribution. In the second step, the update step computes the centroid of each cluster as the mean of the samples in a given cluster, while the maximization step updates the probability and the parameters of each distribution as a weighted average over all the samples. For these reasons, the k -means algorithm is often referred to as a *hard-voting* clustering method, as opposed to the Gaussian mixture model which is referred to as a *soft-voting* clustering method.

The Gaussian mixture model has several advantages over the k -means algorithm.

First, the use of normal distribution densities instead of Euclidean distances dwindles the inflation issue in high-dimensional spaces. Second, the Gaussian mixture model includes covariance matrices, allowing for clusters with elliptical shapes, while the k -means algorithm only includes centroids, forcing clusters to have circular shapes.

Nonetheless, the Gaussian mixture model also has several drawbacks, sharing a few with the k -means algorithm.

First, the number of distributions k is a hyperparameter. Setting a value much different from the actual number of clusters may yield poor clusters. Second, the log-likelihood is not a concave function. Like Lloyd's algorithm, the EM algorithm is guaranteed to converge, but it may converge to a local maximum that is not a global maximum. Several strategies are often applied to address this issue, including sophisticated centroid initialization [22] and running the algorithm numerous times and keeping the best run (i.e., the one yielding the highest log-likelihood). Third, the Gaussian mixture model has more parameters than the k -means algorithm. Therefore, it usually requires more samples to accurately estimate its parameters (in particular the covariance matrices) than the k -means algorithm.

13 Dimensionality Reduction

Dimensionality reduction consists in finding a good mapping from the input space into a space of lower dimension. Dimensionality reduction can either be unsupervised or supervised.

13.1 Principal Component Analysis

For exploratory data analysis, it may be interesting to investigate the variances of the p features and the $\frac{1}{2}p(p - 1)$ covariances or correlations. However, as the value of p increases, this process becomes growingly tedious. Moreover, each feature may explain a small proportion of the total variance. It may be more desirable to have another representation of the data where a small number of features explain most of the total variance, in other words to have a coordinate system adapted to the input data.

Principal component analysis (PCA) consists in finding a representation of the data through *principal components* [25]. The principal components are a sequence of unit vectors such that the i th vector is the best approximation of the data (i.e., maximizing the explained variance) while being orthogonal to the first $i - 1$ vectors.

Figure 23 illustrates principal component analysis when the input space is two-dimensional. On the upper figure, the training data in the original space is plotted. Both features explain about the same amount of the total variance, although one can clearly see that both features are strongly correlated. Principal component analysis identifies a new Cartesian coordinate system based on the input data. On the lower figure, the training data in the new coordinate system is plotted. The first dimension explains much more variance than the second dimension.

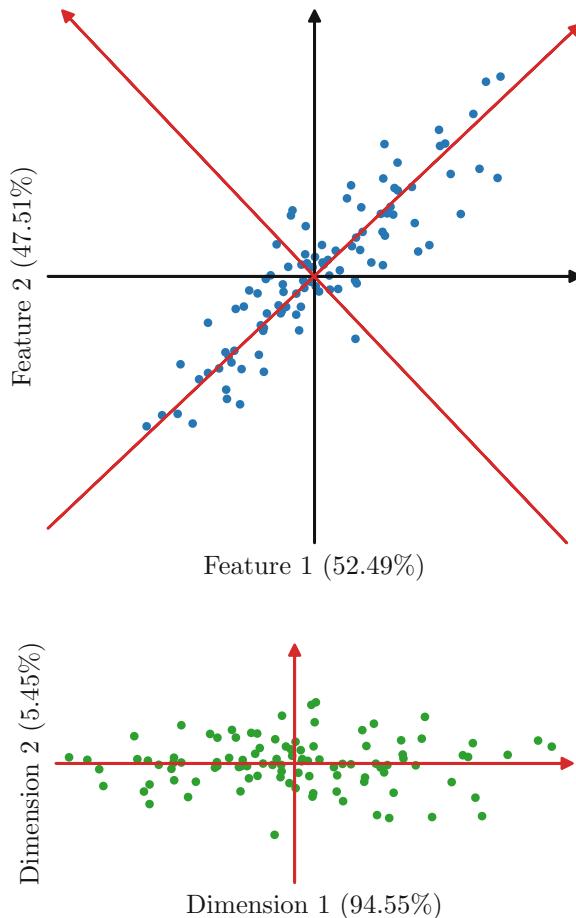


Fig. 23 Illustration of principal component analysis. On the upper figure, the training data in the original space (blue points with black axes) is plotted. Both features explain about the same amount of the total variance, although one can clearly see a linear pattern. Principal component analysis learns a new Cartesian coordinate system based on the input data (red axes). On the lower figure, the training data in the new coordinate system is plotted (green points with red axes). The first dimension explains much more variance than the second dimension

13.1.1 Full Decomposition

Mathematically, given an input matrix $X \in \mathbb{R}^{n \times p}$ that is centered (i.e., the mean value of each column $X_{\cdot j}$ is equal to zero), the objective is to find a matrix $W \in \mathbb{R}^{p \times p}$ such that:

- W is an orthogonal matrix, i.e., its columns are unit vectors and orthogonal to each other.
- The new representation of the input data, denoted by T , consists of the coordinates in the Cartesian coordinate system induced by W (whose columns form an orthogonal basis of \mathbb{R}^p with the Euclidean dot product):

$$T = XW$$

- Each column of W maximizes the explained variance.

Each column $\mathbf{w}_i = \mathbf{W}_{:,i}$ is a principal component. Each input vector \mathbf{x} is transformed into another vector \mathbf{t} using a linear combination of each feature with the weights from the \mathbf{W} matrix:

$$\mathbf{t} = \mathbf{x}^\top \mathbf{W}$$

The first principal component $\mathbf{w}^{(1)}$ is the unit vector that maximizes the explained variance:

$$\begin{aligned}\mathbf{w}_1 &= \arg \max_{\|\mathbf{w}\|=1} \left\{ \sum_{i=1}^n \mathbf{x}^{(i)\top} \mathbf{w} \right\} \\ &= \arg \max_{\|\mathbf{w}\|=1} \{\|\mathbf{X}\mathbf{w}\|\} \\ &= \arg \max_{\|\mathbf{w}\|=1} \{\mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}\} \\ \mathbf{w}_1 &= \arg \max_{\mathbf{w} \in \mathbb{R}^p} \left\{ \frac{\mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \right\}\end{aligned}$$

As $\mathbf{X}^\top \mathbf{X}$ is a positive semi-definite matrix, a well-known result from linear algebra is that $\mathbf{w}^{(1)}$ is the eigenvector associated with the largest eigenvalue of $\mathbf{X}^\top \mathbf{X}$.

The k th component is found by subtracting the first $k-1$ principal components from \mathbf{X} :

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X} \mathbf{w}^{(s)} \mathbf{w}^{(s)\top}$$

and then finding the unit vector that explains the maximum variance from this new data matrix:

$$\mathbf{w}_k = \arg \max_{\|\mathbf{w}\|=1} \{\|\hat{\mathbf{X}}_k \mathbf{w}\|\} = \arg \max_{\mathbf{w} \in \mathbb{R}^p} \left\{ \frac{\mathbf{w}^\top \hat{\mathbf{X}}_k^\top \hat{\mathbf{X}}_k \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \right\}$$

One can show that the eigenvector associated with the k th largest eigenvalue of the $\mathbf{X}^\top \mathbf{X}$ matrix maximizes the quantity to be maximized.

Therefore, the matrix \mathbf{W} is the matrix whose columns are the eigenvectors of the $\mathbf{X}^\top \mathbf{X}$ matrix, sorted by descending order of their associated eigenvalues.

13.1.2 Truncated Decomposition

Since each principal component iteratively maximizes the remaining variance, the first principal components explain most of the total variance, while the last ones explain a tiny proportion of the total variance. Therefore, keeping only a subset of the ordered principal components usually gives a good representation of the input data.

Mathematically, given a number of dimensions l , the new representation is obtained by truncating the matrix of principal components \mathbf{W} to only keep the first l columns, resulting in the submatrix $\mathbf{W}_{:,:l}$.

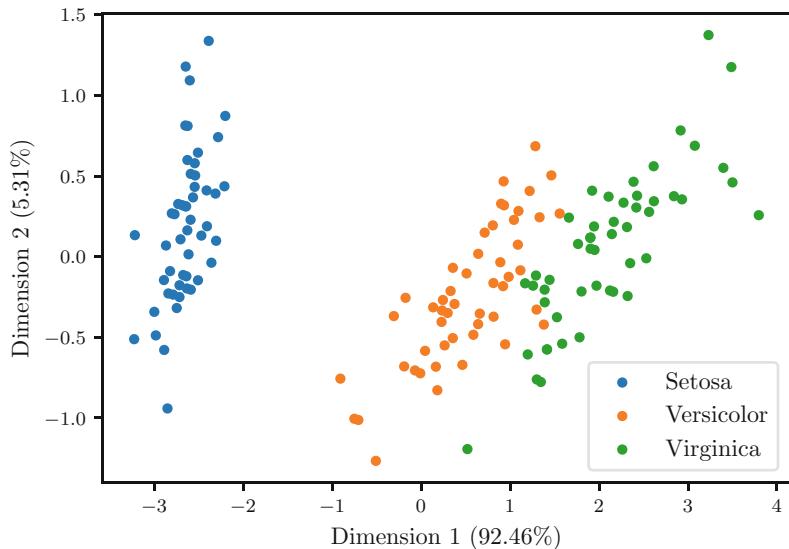


Fig. 24 Illustration of principal component analysis as a dimensionality reduction technique. The Iris flower dataset consists of 50 samples for each of 3 iris species (setosa, versicolor, and virginica) for which 4 features were measured, the length and the width of the sepals and petals, in centimeters. The projection of each sample on the first two principal components is shown in this figure. The first dimension explains most of the variance (92.46%)

$$\tilde{\mathbf{T}} = \mathbf{X} \mathbf{W}_{:,l}$$

Figure 24 illustrates the use of principal component analysis as dimensionality reduction. The Iris flower dataset consists of 50 samples for each of 3 iris species (setosa, versicolor, and virginica) for which 4 features were measured, the length and the width of the sepals and petals, in centimeters. The projection of each sample on the first two principal components is shown in this figure.

13.2 Linear Discriminant Analysis

In Subheading 10, we introduced linear discriminant analysis (LDA) as a classification method. However, it can also be used as a supervised dimensionality reduction method. LDA fits a multivariate normal distribution for each class C_k , so that each class is characterized by its mean vector $\mu_k \in \mathbb{R}^p$ and has the same covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. However, a set of k points lies in a space of dimension at most $k - 1$. For instance, a set of 2 points lies on a line, while a set of 3 points lies on a plane. Therefore, the subspace induced by the k mean vectors μ_k can be used as dimensionality reduction.

There exists another formulation of linear discriminant analysis which is equivalent and more intuitive for dimensionality reduction. Linear discriminant analysis aims to find a linear projection so that the classes are separated as much as possible (i.e., projections of

samples from a same class are close to each other, while projections of samples from different classes are far from each other).

Mathematically, the objective is to find the matrix $\mathbf{W} \in \mathbb{R}^{p \times l}$ (with $l \leq k - 1$) that maximizes the between-class scatter while also minimizing the within-class scatter:

$$\max_{\mathbf{W}} \text{tr}\left(\left(\mathbf{W}^\top \mathbf{S}_w \mathbf{W}\right)^{-1} (\mathbf{W}^\top \mathbf{S}_b \mathbf{W})\right)$$

The within-class scatter matrix \mathbf{S}_w summarizes the diffusion between the mean vector $\boldsymbol{\mu}_k$ of class C_k and all the inputs $\mathbf{x}^{(i)}$ belonging to class C_k , over all the classes:

$$\mathbf{S}_w = \sum_{k=1}^q \sum_{y^{(i)}=C_k} [\mathbf{x}^{(i)} - \boldsymbol{\mu}_k][\mathbf{x}^{(i)} - \boldsymbol{\mu}_k]^\top$$

The between-class scatter matrix \mathbf{S}_b summarizes the diffusion between all the mean vectors:

$$\mathbf{S}_b = \sum_{k=1}^q n_k [\boldsymbol{\mu}_k - \boldsymbol{\mu}] [\boldsymbol{\mu}_k - \boldsymbol{\mu}]^\top$$

where n_k is the proportion of samples belonging to class C_k and $\boldsymbol{\mu} = \sum_{k=1}^q n_k \boldsymbol{\mu}_k = \frac{1}{n} \sum_{i=1}^n \mathbf{x}^{(i)}$ is the mean vector over all the input vectors.

One can show that the \mathbf{W} matrix consists of the first l eigenvectors of the matrix $\mathbf{S}_w^{-1} \mathbf{S}_b$ with the corresponding eigenvalues being sorted in descending order. Just as in principal component analysis, the corresponding eigenvalues can be used to determine the contribution of each dimension. However, the criterion for linear discriminant analysis is different from the one from principal component analysis: it is to maximizing the separability of the classes instead of maximizing the explained variance.

Figure 25 illustrates the use of linear discriminant analysis as a dimensionality reduction technique. We use the same Iris flower dataset as in Fig. 24 illustrating principal component analysis. The projection of each sample on the learned two-dimensional space is shown, and one can see that the first (horizontal) axis is more discriminative of the three classes with linear discriminant analysis than with principal component analysis.

14 Kernel Methods

Kernel methods allow for generalizing linear models to non-linear models with the use of kernel functions.

As mentioned in Subheading 8, the main idea of kernel methods is to first map the input data from the original input space to a feature space and then perform dot products in this feature space.

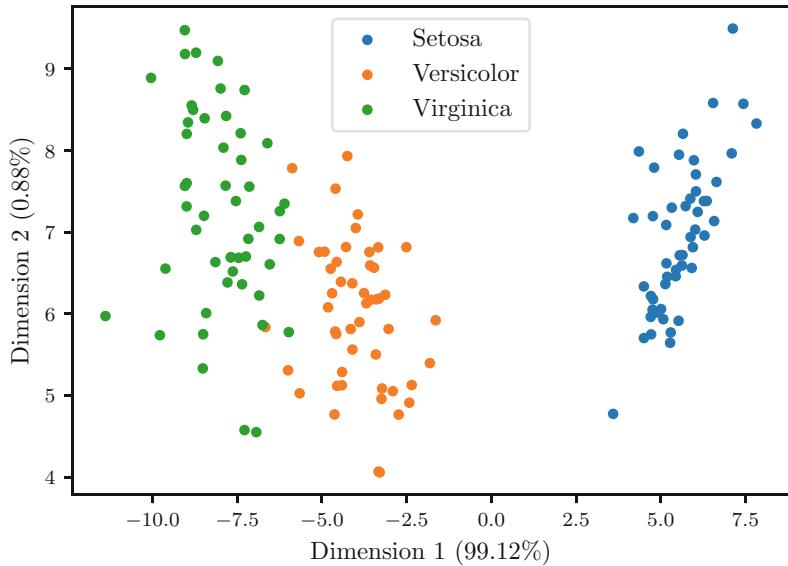


Fig. 25 Illustration of linear discriminant analysis as a dimensionality reduction technique. The Iris flower dataset consists of 50 samples for each of 3 iris species (setosa, versicolor, and virginica) for which 4 features were measured, the length and the width of the sepals and petals, in centimeters. The projection of each sample on the learned two-dimensional space is shown in this figure

Under certain assumptions, an optimal solution of the minimization problem of the cost function admits the following form:

$$f = \sum_{i=1}^n \alpha_i K(\cdot, \mathbf{x}^{(i)})$$

where K is the kernel function which is equal to the dot product in the feature space:

$$\forall \mathbf{x}, \mathbf{x}' \in I, K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}')$$

As this term frequently appears, we denote by \mathbf{K} the $n \times n$ symmetric matrix consisting of the evaluations of the kernel on all the pairs of training samples:

$$\forall i, j \in \{1, \dots, n\}, K_{ij} = K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$$

In this section, we present the extension of two models previously introduced in this chapter, ridge regression and principal component analysis, with kernel functions.

14.1 Kernel Ridge Regression

Kernel ridge regression combines ridge regression with the kernel trick and thus learns a linear function in the space induced by the respective kernel and the training data [2]. For non-linear kernels, this corresponds to a non-linear function in the original input space.

Mathematically, the objective is to find the function f with the following form:

$$f = \sum_{i=1}^n \alpha_i K(\cdot, \mathbf{x}^{(i)})$$

that minimizes the sum of squared errors with a ℓ_2 penalization term:

$$\min_f \sum_{i=1}^n (y^{(i)} - f(\mathbf{x}^{(i)}))^2 + \lambda \|f\|^2$$

The cost function can be simplified using the specific form of the possible functions:

$$\begin{aligned} & \sum_{i=1}^n (y^{(i)} - f(\mathbf{x}^{(i)}))^2 + \lambda \|f\|^2 \\ &= \sum_{i=1}^n \left(y^{(i)} - \sum_{j=1}^n \alpha_j k(\mathbf{x}^{(j)}, \mathbf{x}^{(i)}) \right)^2 + \lambda \left\| \sum_{i=1}^n \alpha_i K(\cdot, \mathbf{x}^{(i)}) \right\|^2 \\ &= \sum_{i=1}^n (y^{(i)} - \boldsymbol{\alpha}^\top \mathbf{K}_{:,i})^2 + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \\ &= \|\mathbf{y} - \mathbf{K} \boldsymbol{\alpha}\|_2^2 + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \end{aligned}$$

Therefore, the minimization problem is:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{K} \boldsymbol{\alpha}\|_2^2 + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$$

for which a solution is given by:

$$\boldsymbol{\alpha}^* = (\mathbf{K} + \lambda I)^{-1} \mathbf{y}$$

Figure 8 illustrates the prediction function of a kernel ridge regression method with a radial basis function kernel. The prediction function is non-linear as the kernel is non-linear.

14.2 Kernel Principal Component Analysis

As mentioned in Subheading 13, principal component analysis consists in finding the linear orthogonal subspace in the original input space such that each principal component explains the most variance. The optimal solution is given by the first eigenvectors of $\mathbf{X}^\top \mathbf{X}$ with the corresponding eigenvalues being sorted in descending order.

With kernel principal component analysis, the objective is to find the linear orthogonal subspace in the feature space such that each principal component in the feature space explains the most variance [26]. The solution is given by the first l eigenvectors $(\boldsymbol{\alpha}_k)_{1 \leq k \leq l}$ of the \mathbf{K} matrix with the corresponding eigenvalues being sorted in descending order. The eigenvectors are normalized in order to be unit vectors in the feature space.

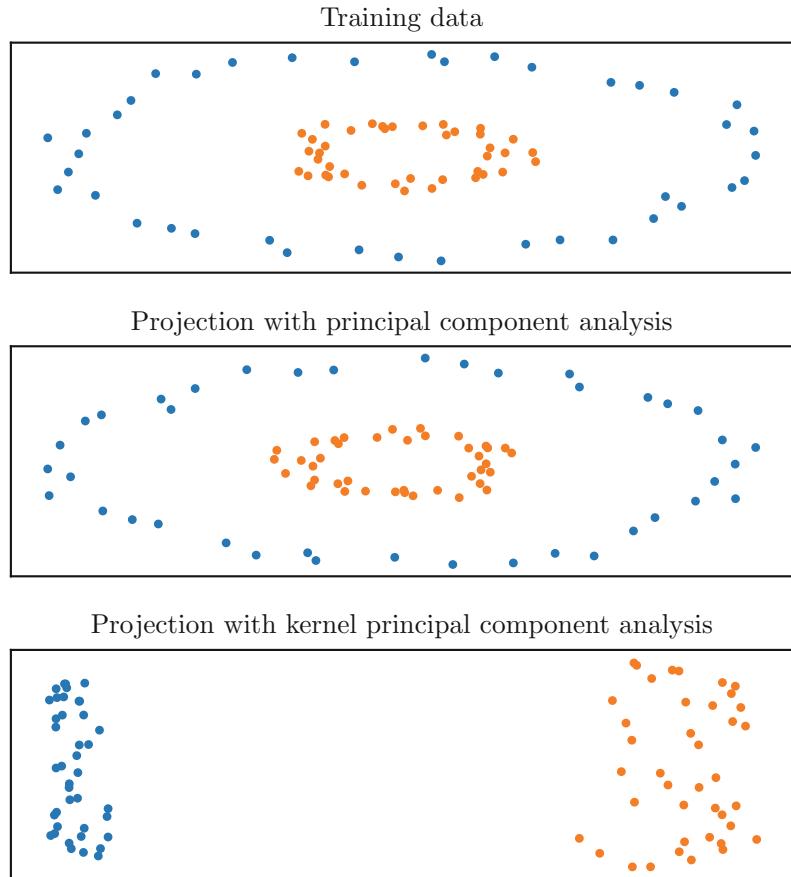


Fig. 26 Illustration of kernel principal component analysis. Some non-linearly separable training data is plotted (top). The projected training data using principal component analysis remains non-linearly separable (middle). The projected training data using kernel principal component analysis (with a non-linear kernel) becomes linearly separable (bottom)

Finally, the projection of any input \mathbf{x} in the original space on the k th component can be computed as:

$$\phi(\mathbf{x})^\top \boldsymbol{\alpha}_k = \sum_{i=1}^n \alpha_{ki} K(\mathbf{x}, \mathbf{x}^{(i)})$$

Figure 26 illustrates the projection of some non-linearly separable classification data with principal component analysis and with kernel principal component analysis with a non-linear kernel. The projected input data becomes linearly separable using kernel principal component analysis, whereas the projected input data using (linear) principal component analysis remains non-linearly separable.

15 Conclusion

In this chapter, we described the main classic machine learning methods. Due to space constraints, the description of some of them was brief. The reader who seeks more details can refer to [5, 6]. All these approaches are implemented in the scikit-learn Python library [27]. A common point of the approaches presented in this chapter is that they use as input a set of given or pre-extracted features. On the contrary, deep learning approaches often provide an end-to-end learning setup within which the features are learned. These techniques are covered in Chaps. 3–6.

Acknowledgements

The authors would like to thank Hicham Janati for his fruitful remarks. The authors would like to acknowledge the extensive documentation of the scikit-learn Python package, in particular its user guide, for the relevant information and references provided. We used the NumPy [28], matplotlib [29], and scikit-learn [27] Python packages to generate all the figures. This work was supported by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6), and by the European Union H2020 program (grant number 826421, project TVB-Cloud).

References

1. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge, MA. <http://www.deeplearningbook.org>
2. Murphy KP (2012) Machine learning: a probabilistic perspective. The MIT Press, Cambridge, MA
3. Bentley JL (1975) Multidimensional binary search trees used for associative searching. Commun ACM 18(9):509–517
4. Omohundro SM (1989) Five balltree construction algorithms. Tech. rep., International Computer Science Institute
5. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin
6. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer series in statistics. Springer, New York
7. Tikhonov AN, Arsenin VY, John F (1977) Solutions of Ill posed problems. Wiley, Washington, New York
8. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Series B (Methodological) 58(1):267–288
9. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Series B (Statistical Methodology) 67(2): 301–320
10. Vapnik VN, Lerner A (1963) Pattern recognition using generalized portrait method. Autom Remote Control 24:774–780
11. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
12. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual

- workshop on computational learning theory. Association for Computing Machinery, Pittsburgh, Pennsylvania, USA, COLT '92, pp 144–152
13. Aizerman MA, Braverman EA, Rozonoer L (1964) Theoretical foundations of the potential function method in pattern recognition learning. In: Automation and remote control, 25, pp 821–837
 14. Schölkopf B, Herbrich R, Smola AJ (2001) A generalized representer theorem. In: Computational learning theory. Springer, Berlin, pp 416–426
 15. Aly M (2005) Survey on multiclass classification methods
 16. James G, Hastie T (1998) The error coding method and PICTs. *J Comput Graph Stat* 7(3):377–387
 17. Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Taylor & Francis, London
 18. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
 19. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Mach Learn* 63(1):3–42
 20. Lloyd S (1982) Least squares quantization in PCM. *IEEE Trans Inform Theory* 28(2): 129–137
 21. Elkan C (2003) Using the triangle inequality to accelerate k-means. In: Proceedings of the twentieth international conference on international conference on machine learning, pp 147–153
 22. Arthur D, Vassilvitskii S (2007) k-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms, pp 1027–1035
 23. Aggarwal CC, Hinneburg A, Keim DA (2001) On the surprising behavior of distance metrics in high dimensional space. In: International conference on database theory. Springer, Berlin, pp 420–434
 24. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B (Methodological)* 39(1):1–38
 25. Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, Berlin
 26. Schölkopf B, Smola AJ, Müller KR (1999) Kernel principal component analysis. In: Advances in kernel methods: support vector learning, MIT Press, Cambridge, MA, pp 327–352
 27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al. (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
 28. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ et al. (2020) Array programming with numpy. *Nature* 585(7825):357–362
 29. Hunter JD (2007) Matplotlib: a 2d graphics environment. *Comput Sci Eng* 9(03):90–95

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made. The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 3

Deep Learning: Basics and Convolutional Neural Networks (CNNs)

Maria Vakalopoulou, Stergios Christodoulidis, Ninon Burgos, Olivier Colliot, and Vincent Lepetit

Abstract

Deep learning belongs to the broader family of machine learning methods and currently provides state-of-the-art performance in a variety of fields, including medical applications. Deep learning architectures can be categorized into different groups depending on their components. However, most of them share similar modules and mathematical formulations. In this chapter, the basic concepts of deep learning will be presented to provide a better understanding of these powerful and broadly used algorithms. The analysis is structured around the main components of deep learning architectures, focusing on convolutional neural networks and autoencoders.

Key words Perceptrons, Backpropagation, Convolutional neural networks, Deep learning, Medical imaging

1 Introduction

Recently, deep learning frameworks have become very popular, attracting a lot of attention from the research community. These frameworks provide machine learning schemes without the need for feature engineering, while at the same time they remain quite flexible. Initially developed for supervised tasks, they are nowadays extended to many other settings. Deep learning, in the strict sense, involves the use of multiple layers of artificial neurons. The first artificial neural networks were developed in the late 1950s with the presentation of the perceptron [1] algorithms. However, limitations related to the computational costs of these algorithms during that period, as well as the often-miscribed claim of Minsky and Papert [2] that perceptrons are not capable of learning non-linear functions such as the XOR, caused a significant decline of interest for further research on these algorithms and contributed to the so-called artificial intelligence winter. In particular, in their book [2], Minsky and Papert discussed that single-layer perceptrons are

only capable of learning linearly separable patterns. It was often incorrectly believed that they also presumed this is the case for multilayer perceptron networks. It took more than 10 years for research on neural networks to recover, and in [3], some of these issues were clarified and further discussed. Even if during this period there was not a lot of research interest for perceptrons, very important algorithms such as the backpropagation algorithm [4–7] and recurrent neural networks [8] were introduced.

After this period, and in the early 2000s, publications by Hinton, Osindero, and Teh [9] indicated efficient ways to train multilayer perceptrons layer by layer, treating each layer as an unsupervised restricted Boltzmann machine and then using supervised backpropagation for the fine-tuning [10]. Such advances in the optimization algorithms and in hardware, in particular graphics processing units (GPUs), increased the computational speed of deep learning systems and made their training easier and faster. Moreover, around 2010, the first large-scale datasets, with ImageNet [11] being one of the most popular, were made available, contributing to the success of deep learning algorithms, allowing the experimental demonstration of their superior performance on several tasks in comparison with other commonly used machine learning algorithms. Finally, another very important factor that contributed to the current popularity of deep learning techniques is their support by publicly available and easy-to-use libraries such as Theano [12], Caffe [13], TensorFlow [14], Keras [15], and PyTorch [16]. Indeed, currently, due to all these publicly available libraries that facilitate collaborative and reproducible research and access to resources from large corporations such as Kaggle, Google Colab, and Amazon Web Services, teaching and research about these algorithms have become much easier.

This chapter will focus on the presentation and discussion of the main components of deep learning algorithms, giving the reader a better understanding of these powerful models. The chapter is meant to be readable by someone with no background in deep learning. The basic notions of machine learning will not be included here; however, the reader should refer to Chap. 2 (reader without a background in engineering or computer science can also refer to Chap. 1 for a lay audience-oriented presentation of these concepts). The rest of this chapter is organized as follows. We will first present the deep feedforward networks focusing on perceptrons, multilayer perceptrons, and the main functions that they are composed of (Subheading 2). Then, we will focus on the optimization of deep neural networks, and in particular, we will formally present the topics of gradient descent, backpropagation, as well as the notions of generalization and overfitting (Subheading 3). Subheading 4 will focus on convolutional neural networks discussing in detail the basic convolution operations, while Subheading 5 will give an overview of the autoencoder architectures.

2 Deep Feedforward Networks

In this section, we will present the early deep learning approaches together with the main functions that are commonly used in deep feedforward networks. Deep feedforward networks are a set of parametric, non-linear, and hierarchical representation models that are optimized with stochastic gradient descent. In this definition, the term parametric holds due to the parameters that we need to learn during the training of these models, the non-linearity due to the non-linear functions that they are composed of, and the hierarchical representation due to the fact that the output of one function is used as the input of the next in a hierarchical way.

2.1 Perceptrons

The perceptron [1] was originally developed for supervised binary classification problems, and it was inspired by works from neuroscientists such as Donald Hebb [17]. It was built around a non-linear neuron, namely, the McCulloch-Pitts model of a neuron. More formally, we are looking for a function $f(\mathbf{x}; \mathbf{w}, b)$ such that $f(\cdot; \mathbf{w}, b) : \mathbf{x} \in \mathbb{R}^p \rightarrow \{+1, -1\}$ where \mathbf{w} and b are the parameters of f and the vector $\mathbf{x} = [x_1, \dots, x_p]^T$ is the input. The training set is $\{(\mathbf{x}^{(i)}, y^{(i)})\}$. In particular, the perceptron relies on a linear model for performing the classification:

$$f(\mathbf{x}; \mathbf{w}, b) = \begin{cases} +1 & \text{if } \mathbf{w}^T \mathbf{x} + b \geq 0 \\ -1 & \text{otherwise} \end{cases}. \quad (1)$$

Such a model can be interpreted geometrically as a hyperplane that can appropriately divide data points that are linearly separable. Moreover, one can observe that, in the previous definition, a perceptron is a combination of a weighted summation between the elements of the input vector \mathbf{x} combined with a step function that performs the decision for the classification. Without loss of generality, this step function can be replaced by other activation functions such as the sigmoid, hyperbolic tangent, or softmax functions (see Subheading 2.3); the output simply needs to be thresholded to assign the $+1$ or -1 class. Graphically, a perceptron is presented in Fig. 1 on which each of the elements of the input is described as a neuron and all the elements are combined by weighting with the models' parameters and then passed to an activation function for the final decision.

During the training process and similarly to the other machine learning algorithms, we need to find the optimal parameters \mathbf{w} and b for the perceptron model. One of the main innovations of Rosenblatt was the proposition of the learning algorithm using an iterative process. First, the weights are initialized randomly, and then using one sample $(\mathbf{x}^{(i)}, y^{(i)})$ of the training set, the prediction of the

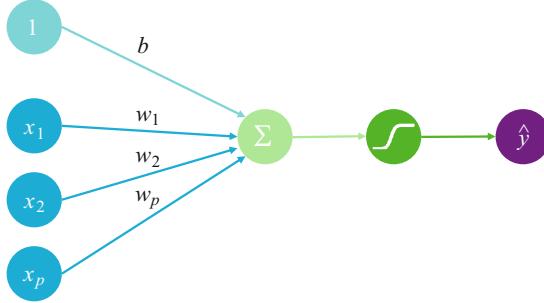


Fig. 1 A simple perceptron model. The input elements are described as neurons and combined for the final prediction \hat{y} . The final prediction is composed of a weighted sum and an activation function

perceptron is calculated. If the prediction is correct, no further action is needed, and the next data point is processed. If the prediction is wrong, the weights are updated with the following rule: the weights are increased in case the prediction is smaller than the ground-truth label $y^{(i)}$ and decreased if the prediction is higher than the ground-truth label. This process is repeated until no further errors are made for the data points. A pseudocode of the training or convergence algorithm is presented in Algorithm 1 (note that in this version, it is assumed that the data is linearly separable).

Algorithm 1 Train perceptron

```

procedure TRAIN( $\{(\mathbf{x}^{(i)}, y^{(i)})\}$ )
    Initialization: initialize randomly the weights  $\mathbf{w}$  and bias  $b$ 
    while  $\exists i \in \{1, \dots, n\}, f(\mathbf{x}^{(i)}; \mathbf{w}, b) \neq y^{(i)}$  do
        Pick  $i$  randomly
        error =  $y^{(i)} - f(\mathbf{x}^{(i)}; \mathbf{w}, b)$ 
        if error  $\neq 0$  then
             $\mathbf{w} \leftarrow \mathbf{w} + error \cdot \mathbf{x}^{(i)}$ 
             $b \leftarrow b + error$ 

```

Originally, the perceptron has been proposed for binary classification tasks. However, this algorithm can be generalized for the case of multiclass classification, $f_c(\mathbf{x}; \mathbf{w}, b)$, where $c \in \{1, \dots, C\}$ are the different classes. This can be easily achieved by adding more neurons to the output layer of the perceptron. That way, the number of output neurons would be the same as the number of possible outputs we need to predict for the specific problem. Then, the final decision can be made by choosing the maximum of the different output neurons $f_n = \max_{c \in \{1, \dots, C\}} f_c(\mathbf{x}; \mathbf{w}, b)$.

Finally, in the following, we will integrate the bias b in the weights \mathbf{w} (and thus add 1 as the first element of the input vector $\mathbf{x} = [1, x_1, \dots, x_p]^\top$). The model can then be rewritten as $f(\mathbf{x}; \mathbf{w})$ such that $f(\cdot; \mathbf{w}) : \mathbf{x} \in \mathbb{R}^{p+1} \rightarrow \{+1, -1\}$.

2.2 Multilayer Perceptrons

The limitation of perceptrons to linear problems can be overcome by using multilayer perceptions, often denoted as MLP. An MLP consists of at least three layers of neurons: the input layer, a hidden layer, and an output layer. Except for the input neurons, each neuron uses a non-linear activation function, making it capable of distinguishing data that is not linearly separable. These layers can also be called fully connected layers since they connect all the neurons of the previous and of the current layer. It is absolutely crucial to keep in mind that non-linear functions are necessary for the network to find non-linear separations in the data (otherwise, all the layers could simply be collapsed together into a single gigantic linear function).

2.2.1 A Simple Multilayer Network

Without loss of generality, an MLP with one hidden layer can be defined as:

$$\begin{cases} \mathbf{z}(\mathbf{x}) = g(\mathbf{W}^1 \mathbf{x}) \\ \hat{\mathbf{y}} = f(\mathbf{x}; \mathbf{W}^1, \mathbf{W}^2) = \mathbf{W}^2 \mathbf{z}(\mathbf{x}) \end{cases}, \quad (2)$$

where $g(x) : \mathbb{R} \rightarrow \mathbb{R}$ denotes the non-linear function (which can be applied element-wise to a vector), \mathbf{W}^1 the matrix of coefficients of the first layer, and \mathbf{W}^2 the matrix of coefficients of the second layer.

Equivalently, one can write:

$$\hat{y}_c = \sum_{j=1}^{d_1} \mathbf{W}_{(c,j)}^2 g(\mathbf{W}_{(j)}^{1\top} \mathbf{x}), \quad (3)$$

where d_1 is the number of neurons for the hidden layer which defines the width of the network, $\mathbf{W}_{(j)}^1$ denotes the first column of the matrix \mathbf{W}^1 , and $\mathbf{W}_{(c,j)}^2$ denotes the c, j element of the matrix \mathbf{W}^2 . Graphically, a two-layer perceptron is presented in Fig. 2 on

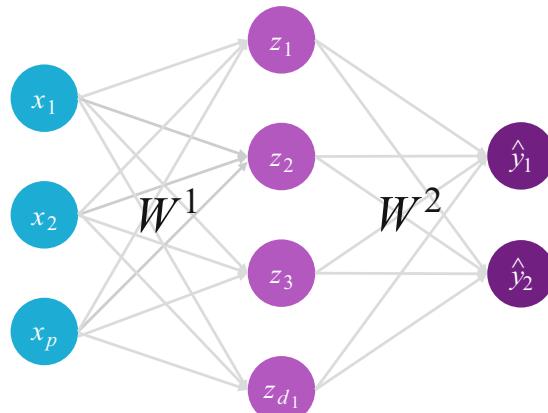


Fig. 2 An example of a simple multilayer perceptron model. The input layer is fed into a hidden layer (\mathbf{z}), which is then combined for the last output layer providing the final prediction

which the input neurons are fed into a hidden layer whose neurons are combined for the final prediction.

There were a lot of research works indicating the capacity of feedforward neural networks with a single hidden layer of finite size to approximate continuous functions. In the late 1980s, the first proof was published [18] for sigmoid activation functions (*see* Subheading 2.3 for the definition) and was generalized to other functions for feedforward multilayer architectures [19–21]. In particular, these works prove that any continuous function can be approximated under mild conditions as closely as wanted by a three-layer network. As $N \rightarrow \infty$, any continuous function f can be approximated by some neural network \hat{f} , because each component $\mathcal{G}(\mathbf{W}_{(j)}^T \mathbf{x})$ behaves like a basis function and functions in a suitable space admit a basis expansion. However, since N may need to be very large, introducing some limitations for these types of networks, deeper networks, with more than one hidden layer, can provide good alternatives.

2.2.2 Deep Neural Network

The simple MLP networks can be generalized to deeper networks with more than one hidden layer that progressively generate higher-level features from the raw input. Such networks can be written as:

$$\begin{cases} z_1(\mathbf{x}) = \mathcal{G}(\mathbf{W}^1 \mathbf{x}) \\ \dots \\ z_k(\mathbf{x}) = \mathcal{G}(\mathbf{W}^k z_{k-1}(\mathbf{x})) \\ \dots \\ \hat{y} = f(\mathbf{x}; \mathbf{W}^1, \dots, \mathbf{W}^K) = z_K(z_{K-1}(\dots(z_1(\mathbf{x})))) \end{cases}, \quad (4)$$

where K denotes the number of layers for the neural network, which defines the depth of the network. In Fig. 3, a graphical representation of the deep multilayer perceptron is presented. Once again, the input layer is fed into the different hidden layers of the network in a hierarchical way such that the output of one layer is the input of the next one. The last layer of the network corresponds to the output layer, which makes the final prediction of the model.

As for networks with one hidden layer, they are also universal approximators. However, the approximation theory for deep networks is less understood compared with neural networks with one hidden layer. Overall, deep neural networks excel at representing the composition of functions.

So far, we have described neural networks as simple chains of layers, applied in a hierarchical way, with the main considerations being the depth of the network (the number of layers K) and the

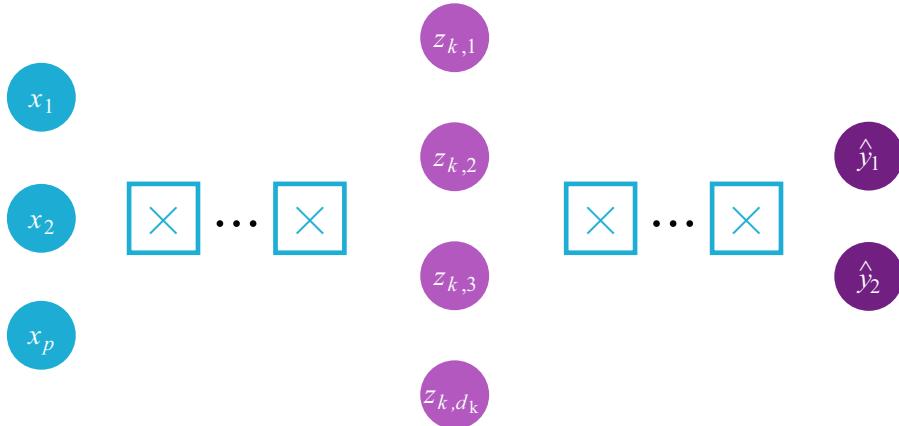


Fig. 3 An example of a deep neural network. The input layer, the k th layer of the deep neural network, and the output layer are presented in the figure

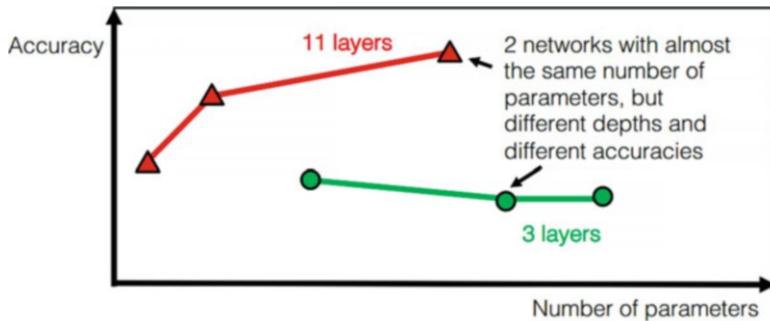


Fig. 4 Comparison of two different networks with almost the same number of parameters, but different depths. Figure inspired by Goodfellow et al. [24]

width of each k layer (the number of neurons d_k). Overall, there are no rules for the choice of the K and d_k parameters that define the architecture of the MLP. However, it has been shown empirically that deeper models perform better. In Fig. 4, an overview of 2 different networks with 3 and 11 hidden layers is presented with respect to the number of parameters and their accuracy. For each architecture, the number of parameters varies by changing the number of neurons d_k . One can observe that, empirically, deeper networks achieve better performance using approximately the same or a lower number of parameters. Additional evidence to support these empirical findings is a very active field of research [22, 23].

Neural networks can come in a variety of models and architectures. The choice of the proper architecture and type of neural network depends on the type of application and the type of data.

Most of the time, the best architecture is defined empirically. In the next section, we will discuss the main functions used in neural networks.

2.3 Main Functions

A neural network is a composition of different functions also called modules. Most of the times, these functions are applied in a sequential way. However, in more complicated designs (e.g., deep residual networks), different ways of combining them can be designed. In the following subsections, we will discuss the most commonly used functions that are the backbones of most perceptrons and multi-layer perceptron architectures. One should note, however, that a variety of functions can be proposed and used for different deep learning architectures with the constraint to be differentiable – almost – everywhere. This is mainly due to the way that deep neural networks are trained, and this will be discussed later in the chapter.

2.3.1 Linear Functions

One of the most fundamental functions used in deep neural networks is the simple linear function. Linear functions produce a linear combination of all the nodes of one layer of the network, weighted with the parameters \mathbf{W} . The output signal of the linear function is $\mathbf{W}\mathbf{x}$, which is a polynomial of degree one. While it is easy to solve linear equations, they have less power to learn complex functional mappings from data. Moreover, when the number of samples is much larger than the dimension of the input space, the probability that the data is linearly separable comes close to zero (Box 1). This is why they need to be combined with non-linear functions, also called activation functions (the name activation has been initially inspired by biology as the neuron will be active or not depending on the output of the function).

Box 1: Function Counting Theorem

The so-called Function Counting Theorem (Cover [25]) counts the number of linearly separable dichotomies of n points in general position in \mathbb{R}^p . The theorem shows that, out of the total 2^n dichotomies, only $C(n, p) = 2 \sum_{j=0}^p \binom{n-1}{j}$ are homogeneously, linearly separable.

When $n \gg p$, the probability of a dichotomy to be linearly separable converges to zero. This indicates the need for the integration of non-linear functions into our modeling and architecture design. Note that $n \gg p$ is a typical regime in machine learning and deep learning applications where the number of samples is very large.

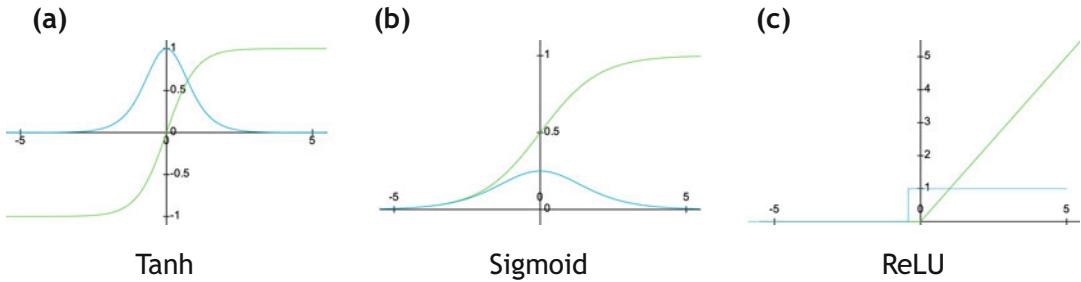


Fig. 5 Overview of different non-linear functions (in green) and their first-order derivative (blue). (a) Hyperbolic tangent function (tanh), (b) sigmoid, and (c) rectified linear unit (ReLU)

2.3.2 Non-linear Functions

Hyperbolic Tangent Function (tanh)

One of the most important components of deep neural networks is the non-linear functions, also called activation functions. They convert the linear input signal of a node into non-linear outputs to facilitate the learning of high-order polynomials. There are a lot of different non-linear functions in the literature. In this subsection, we will discuss the most classical non-linearities.

One of the most standard non-linear functions is the hyperbolic tangent function, aka the tanh function. Tanh is symmetric around the origin with a range of values varying from -1 to 1 . The biggest advantage of the tanh function is that it produces a zero-centered output (Fig. 5a), thereby supporting the backpropagation process that we will cover in the next section. The tanh function is used extensively for the training of multilayer neural networks. Formally, the tanh function, together with its gradient, is defined as:

$$\begin{aligned} g &= \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \\ \frac{\partial g}{\partial x} &= 1 - \tanh^2(x) \end{aligned} \quad . \quad (5)$$

One of the downsides of tanh is the saturation of gradients that occurs for large or small inputs. This can slow down the training of the networks.

Sigmoid

Similar to tanh, the sigmoid is one of the first non-linear functions that were used to compose deep learning architectures. One of the main advantages is that it has a range of values varying from 0 to 1 (Fig. 5b) and therefore is especially used for models that aim to predict a probability as an output. Formally, the sigmoid function, together with its gradient, is defined as:

$$\begin{aligned} g &= \sigma(x) = \frac{1}{1 + e^{-x}} \\ \frac{\partial g}{\partial x} &= \sigma(x)(1 - \sigma(x)) \end{aligned} \quad . \quad (6)$$

Note that this is in fact the logistic function, which is a special case of the more general class of sigmoid function. As it is indicated in Fig. 5b, the sigmoid gradient vanishes for large or small inputs making the training process difficult. However, in case it is used for the output units which are not latent variables and on which we have access to the ground-truth labels, sigmoid may be a good option.

Rectified Linear Unit (ReLU)

ReLU is considered among the default choice of non-linearity. Some of the main advantages of ReLU include its efficient calculation and better gradient propagation with fewer vanishing gradient problems compared to the previous two activation functions [26]. Formally, the ReLU function, together with its gradient, is defined as:

$$\begin{aligned} g &= \max(0, x) \\ \frac{\partial g}{\partial x} &= \begin{cases} 0, & \text{if } x \leq 0 \\ 1, & \text{if } x > 0 \end{cases}. \end{aligned} \quad (7)$$

As it is indicated in Fig. 5c, ReLU is differentiable anywhere else than zero. However, this is not a very important problem as the value of the derivative at zero can be arbitrarily chosen to be 0 or 1. In [27], the authors empirically demonstrated that the number of iterations required to reach 25% training error on the CIFAR-10 dataset for a four-layer convolutional network was six times faster with ReLU than with tanh neurons. On the other hand, and as discussed in [28], ReLU-type neural networks which yield a piecewise linear classifier function produce almost always high confidence predictions far away from the training data. However, due to its efficiency and popularity, many variations of ReLU have been proposed in the literature, such as the leaky ReLU [29] or the parametric ReLU [30]. These two variations both address the problem of dying neurons, where some ReLU neurons die for all inputs and remain inactive no matter what input is supplied. In such a case, no gradient flows from these neurons, and the training of the neural network architecture is affected. Leaky ReLU and parametric ReLU change the $g(x)=0$ part, by adding a slope and extending the range of ReLU.

Swish

The choice of the activation function in neural networks is not always easy and can greatly affect performance. In [31], the authors performed a combination of exhaustive and reinforcement learning-based searches to discover novel activation functions. Their experiments discovered a new activation function that is called Swish and is defined as:

$$\begin{aligned} \mathcal{J} &= \mathbf{x} \cdot \sigma(\beta \mathbf{x}) \\ \frac{\partial \mathcal{J}}{\partial \mathbf{x}} &= \beta \mathcal{J}(\mathbf{x}) + \sigma(\beta \mathbf{x})(1 - \beta \mathcal{J}(\mathbf{x})) \end{aligned}, \quad (8)$$

where σ is the sigmoid function and β is either a constant or a trainable parameter. Swish tends to work better than ReLU on deeper models, as it has been shown experimentally in [31] in different domains.

Softmax

Softmax is often used as the last activation function of a neural network. In practice, it normalizes the output of a network to a probability distribution over the predicted output classes. Softmax is defined as:

$$\text{Softmax}(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i}}{\sum_j^C e^{\mathbf{x}_j}}. \quad (9)$$

The softmax function takes as input a vector \mathbf{x} of C real numbers and normalizes it into a probability distribution consisting of C probabilities proportional to the exponentials of the input numbers. However, a limitation of softmax is that it assumes that every input \mathbf{x} belongs to at least one of the C classes (which is not the case in practice, i.e., the network could be applied to an input that does not belong to any of the classes).

2.3.3 Loss Functions

Besides the activation functions, the loss function (which defines the cost function) is one of the main elements of neural networks. It is the function that represents the error for a given prediction. To that purpose, for a given training sample, it compares the prediction $f(\mathbf{x}^{(i)}; \mathbf{W})$ to the ground truth $y^{(i)}$ (here we denote for simplicity as \mathbf{W} all the parameters of the network, combining all the $\mathbf{W}^1, \dots, \mathbf{W}^K$ in the multilayer perceptron shown above). The loss is denoted as $\ell(y, f(\mathbf{x}; \mathbf{W}))$. The average loss across the n training samples is called the cost function and is defined as:

$$J(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)}, f(\mathbf{x}^{(i)}; \mathbf{W})), \quad (10)$$

where $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1..n}$ composes the training set. The aim of the training will be to find the parameters \mathbf{W} such that $J(\mathbf{W})$ is minimized. Note that, in deep learning, one often calls the cost function the loss function, although, strictly speaking, the loss is for a given sample, and the cost is averaged across samples. Besides, the objective function is the overall function to minimize, including the cost and possible regularization terms. However, in the remainder of this chapter, in accordance with common usage in deep learning, we will sometimes use the term loss function instead of cost function.

In neural networks, the loss function can be virtually any function that is differentiable. Below we present the two most common losses, which are, respectively, used for classification or regression problems. However, specific losses exist for other tasks, such as segmentation, which are covered in the corresponding chapters.

Cross-Entropy Loss

One of the most basic loss functions for classification problems corresponds to the cross-entropy between the expected values and the predicted ones. It leads to the following cost function:

$$J(\mathbf{W}) = - \sum_{i=1}^n \log(P(y=y^{(i)}|\mathbf{x}=\mathbf{x}^{(i)}; \mathbf{W})), \quad (11)$$

where $P(y=y^{(i)}|\mathbf{x}=\mathbf{x}^{(i)}; \mathbf{W})$ is the probability that a given sample is correctly classified.

The cross-entropy can also be seen here as the negative log-likelihood of the training set given the predictions of the network. In other words, minimizing this loss function corresponds to maximizing the likelihood:

$$J(\mathbf{W}) = \prod_{i=1}^n P(y=y^{(i)}|\mathbf{x}=\mathbf{x}^{(i)}; \mathbf{W}). \quad (12)$$

Mean Squared Error Loss

For regression problems, the mean squared error is one of the most basic cost functions, measuring the average of the squares of the errors, which is the average squared difference between the predicted values and the real ones. The mean squared error is defined as:

$$J(\mathbf{W}) = \sum_{i=1}^n \|y^{(i)} - f(\mathbf{x}^{(i)}; \mathbf{W})\|^2. \quad (13)$$

3 Optimization of Deep Neural Networks

Optimization is one of the most important components of neural networks, and it focuses on finding the parameters \mathbf{W} that minimize the loss function $J(\mathbf{W})$. Overall, optimization is a difficult task. Traditionally, the optimization process is performed by carefully designing the loss function and integrating its constraints to ensure that the optimization process is convex (and thus, one can be sure to find the global minimum). However, neural networks are non-convex models, making their optimization challenging, and, in general, one does not find the global minimum but only a local one. In the next sections, the main components of their optimization will be presented, giving a general overview of the optimization process, its challenges, and common practices.

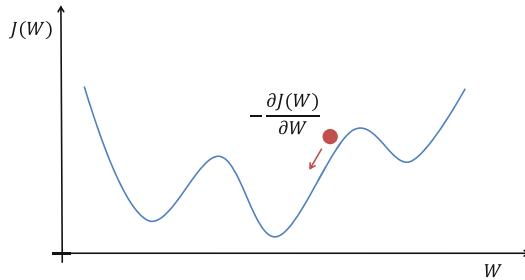


Fig. 6 The gradient descent algorithm. This first-order optimization algorithm is finding a local minimum by taking steps toward the opposite direction of the gradient

3.1 Gradient Descent

Gradient descent is an iterative optimization algorithm that is among the most popular and basic algorithms in machine learning. It is a first-order¹ optimization algorithm, which is finding a local minimum of a differentiable function. The main idea of gradient descent is to take iterative steps toward the opposite direction of the gradient of the function that needs to be optimized (Fig. 6).

That way, the parameters W of the model are updated by:

$$W^{t+1} \leftarrow W^t - \eta \frac{\partial J(W^t)}{\partial W^t}, \quad (14)$$

where t is the iteration and η , called learning rate, is the hyperparameter that indicates the magnitude of the step that the algorithm will take.

Besides its simplicity, gradient descent is one of the most commonly used algorithms. More sophisticated algorithms require computing the Hessian (or an approximation) and/or its inverse (or an approximation). Even if these variations could give better optimization guarantees, they are often more computationally expensive, making gradient descent the default method for optimization.

In the case of convex functions, the optimization problem can be reduced to the problem of finding a local minimum. Any local minimum is then guaranteed to be a global minimum, and gradient descent can identify it. However, when dealing with non-convex functions, such as neural networks, it is possible to have many local minima making the use of gradient descent challenging. Neural networks are, in general, non-identifiable [24]. A model is said to be identifiable if it is theoretically possible, given a sufficiently large training set, to rule out all but one set of the model's parameters. Models with latent variables, such as the hidden layers of neural networks, are often not identifiable because we can obtain equivalent models by exchanging latent variables with each other.

¹ First-order means here that the first-order derivatives of the cost function are used as opposed to second-order algorithms that, for instance, use the Hessian.

However, all these minima are often almost equivalent to each other in cost function value. In that case, these local minima are not a problematic form of non-convexity. It remains an open question whether there exist many local minima with a high cost that prevent adequate training of neural networks. However, it is currently believed that most local minima, at least as found by modern optimization procedures, will correspond to a low cost (even though not to identical costs) [24].

For \mathbf{W}^* to be a local minimum, we need mainly two conditions to be fulfilled:

- $\left\| \frac{\partial J}{\partial \mathbf{W}}(\mathbf{W}^*) \right\| = 0$.
- All the eigenvalues of $\left(\frac{\partial^2 J}{\partial \mathbf{W}^2}(\mathbf{W}^*) \right)$ to be positive.

For random functions in n dimensions, the probability for the eigenvalues to be all positive is $\frac{1}{n}$. On the other hand, the ratio of the number of saddle points to local minima increases exponentially with n [32]. A saddle point, or critical point, is a point where the derivatives are zero without being a minimum of the function. Such points could result in a high error making the optimization with gradient descent challenging. In [32], this issue is discussed, and an optimization algorithm that leverages second-order curvature information is proposed to deal with this issue for deep and recurrent networks.

3.1.1 Stochastic Gradient Descent

Gradient descent efficiency is not enough when it comes to machine learning problems with large numbers of training samples. Indeed, this is the case for neural networks and deep learning which often rely on hundreds or thousands of training samples. Updating the parameters \mathbf{W} after calculating the gradient using all the training samples would lead to a tremendous computational complexity of the underlying optimization algorithm [33]. To deal with this problem, the stochastic gradient descent (SGD) algorithm is a drastic simplification. Instead of computing the $\frac{\partial J(\mathbf{W})}{\partial \mathbf{W}}$ exactly, each iteration estimates this gradient on the basis of a small set of randomly picked examples, as follows:

$$\mathbf{W}^{t+1} \leftarrow \mathbf{W}^t - \eta_t G(\mathbf{W}^t), \quad (15)$$

where

$$G(\mathbf{W}^t) = \frac{1}{K} \sum_{k=1}^K \frac{\partial J_{(i_k)}}{\partial \mathbf{W}} \mathbf{W}^t, \quad (16)$$

where J_{i_k} is the loss function at training sample i_k , $\{(\mathbf{x}^{(i_k)}, y^{(i_k)})\}_{k=1 \dots K}$ is the small subset of K training samples ($K \ll N$). This subset of K samples is called a mini-batch or sometimes a batch.² In such a way, the iteration cost of stochastic

² Note that, as often in deep learning, the terminology can be confusing. In isolation, the term batch is usually a synonym of mini-batch. On the contrary, batch gradient descent means computing the gradient using all training samples and not only a mini-batch [24].

gradient descent will be $\mathcal{O}(K)$ and for gradient descent $\mathcal{O}(N)$. The ideal choice for the batch size is a debated question. First, an upper limit for the batch size is often simply given by the available GPU memory, in particular when the size of the input data is large (e.g., 3D medical images). Besides, choosing K as a power of 2 often leads to more efficient computations. Finally, small batch sizes tend to have a regularizing effect which can be beneficial [24]. In any case, the ideal batch size usually depends on the application, and it is not uncommon to try different batch sizes. Finally, one calls an epoch a complete pass over the whole training set (meaning that each training sample has been used once). The number of epochs is the number of full passes over the whole training set. It should not be confused with the number of iterations which is the number of mini-batches that have been processed.

Note that various improvements over traditional SGD have been introduced, leading to more efficient optimization methods. These state-of-the-art optimization methods are presented in Subheading 3.4.

Box 2: Convergence of SGD Theorem

In [34], the authors prove that stochastic gradient descent converges if the network is sufficiently overparametrized. Let $(\mathbf{x}^{(i)}, y^{(i)})_{1 \leq i \leq n}$ be a training set satisfying $\min_{i,j: i \neq j} \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|_2 > \delta > 0$. Consider fitting the data using a feedforward neural network with ReLU activations. Denote by D (resp. W) the depth (resp. width) of the network. Suppose that the neural network is sufficiently overparametrized, i.e.:

$$W \gg \text{polynomial}\left(n, D, \frac{1}{\delta}\right). \quad (17)$$

Then, with high probability, running SGD with *some random initialization* and properly chosen step sizes η_t yields $J(\mathbf{W}^t) < \epsilon$ in $t \propto \log \frac{1}{\epsilon}$.

3.2 Backpropagation

The training of neural networks is performed with backpropagation. Backpropagation computes the gradient of the loss function with respect to the parameters of the network in an efficient and local way. This algorithm was originally introduced in 1970. However, it started becoming very popular after the publication of [6], which indicated that backpropagation works faster than other methods that had been proposed back then for the training of neural networks.



Fig. 7 A multilayer perceptron with one hidden layer

The backpropagation algorithm works by computing the gradient of the loss function (J) with respect to each weight by the chain rule, computing the gradient one layer at a time, and iterating backward from the last layer to avoid redundant calculations of intermediate terms in the chain rule. In Fig. 7, an example of a multilayer perceptron with one hidden layer is presented. In such a network, the backpropagation is calculated as:

$$\begin{aligned}\frac{\partial J(\mathbf{W})}{\partial w_2} &= \frac{\partial J(\mathbf{W})}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_2} \\ \frac{\partial J(\mathbf{W})}{\partial w_1} &= \frac{\partial J(\mathbf{W})}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_1} = \frac{\partial J(\mathbf{W})}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_1} \times \frac{\partial z_1}{\partial w_1}.\end{aligned}\quad (18)$$

Overall, backpropagation is very simple and local. However, the reason why we can train a highly non-convex machine with many local minima, like neural networks, with a strong local learning algorithm is not really known even today. In practice, backpropagation can be computed in different ways, including manual calculation, numerical differentiation using finite difference approximation, and symbolic differentiation. Nowadays, deep learning frameworks such as [14, 16] use automatic differentiation [35] for the application of backpropagation.

3.3 Generalization and Overfitting

Similar to all the machine learning algorithms (discussed in Chapter 2), neural networks can suffer from poor generalization and overfitting. These problems are caused mainly by the optimization of the parameters of the models performed in the $\{(\mathbf{x}_i, y_i)\}_{i=1, \dots, n}$ training set, while we need the model to perform well on other unseen data that are not available during the training. More formally, in the case of cross-entropy, the loss that we would like to minimize is:

$$J(\mathbf{W}) = -\log \prod_{(\mathbf{x}, y) \in T_T} P(y = y | \mathbf{x} = \mathbf{x}; \mathbf{W}), \quad (19)$$

where T_T is the set of any data, not available during training. In practice, a small validation set T_V is used to evaluate the loss on unseen data. Of course, this validation set should be distinct from the training set. It is extremely important to keep in mind that the performance obtained on the validation set is generally biased upward because the validation set was used to perform early stopping or to choose regularization parameters. Therefore, one should have an independent test set that has been isolated at the

beginning, has not been used in any way during training, and is only used to report the performance (see Chap. 20 for details). In case one cannot have an additional independent test set due to a lack of data, one should be aware that the performance may be biased and that this is a limitation of the specific study.

To avoid overfitting and improve the generalization performance of the model, usually, the validation set is used to monitor the loss during the training of the networks. Tracking the training and validation losses over the number of epochs is essential and provides important insights into the training process and the selected hyperparameters (e.g., choice of learning rate). Recent visualization tools such as TensorBoard³ or Weights & Biases⁴ make this tracking easy. In the following, we will also mention some of the most commonly applied optimization techniques that help with preventing overfitting.

Early Stopping Using the reported training and validation errors, the best model in terms of performance and generalization power is selected. In particular, early stopping, which corresponds to selecting a model corresponding to an earlier time point than the final epoch, is a common way to prevent overfitting [36]. Early stopping is a form of regularization for models that are trained with an iterative method, such as gradient descent and its variants. Early stopping can be implemented with different criteria. However, generally, it requires the monitoring of the performance of the model on a validation set, and the model is selected when its performance degrades or its loss increases. Overall, early stopping should be used almost universally for the training of neural networks [24]. The concept of early stopping is illustrated in Fig. 8.

Weight Regularization Similar to other machine learning methods (Chap. 2), weight regularization is also a very commonly used technique for avoiding overfitting in neural networks. More specifically, during the training of the model, the weights of the network start growing in size in order to specialize the model to the training data. However, large weights tend to cause sharp transitions in the different layers of the network and, that way, large changes in the output for only small changes in the inputs [37]. To handle this problem, during the training process, the weights can be updated in such a way that they are encouraged to be small, by adding a penalty to the loss function, for instance, the ℓ_2 norm of the parameters $\lambda \|\mathbf{W}\|^2$, where λ is a trade-off parameter between the loss and the regularization. Since weight regularization is quite popular in

³ <https://www.tensorflow.org/tensorboard>.

⁴ <https://wandb.ai/site>.

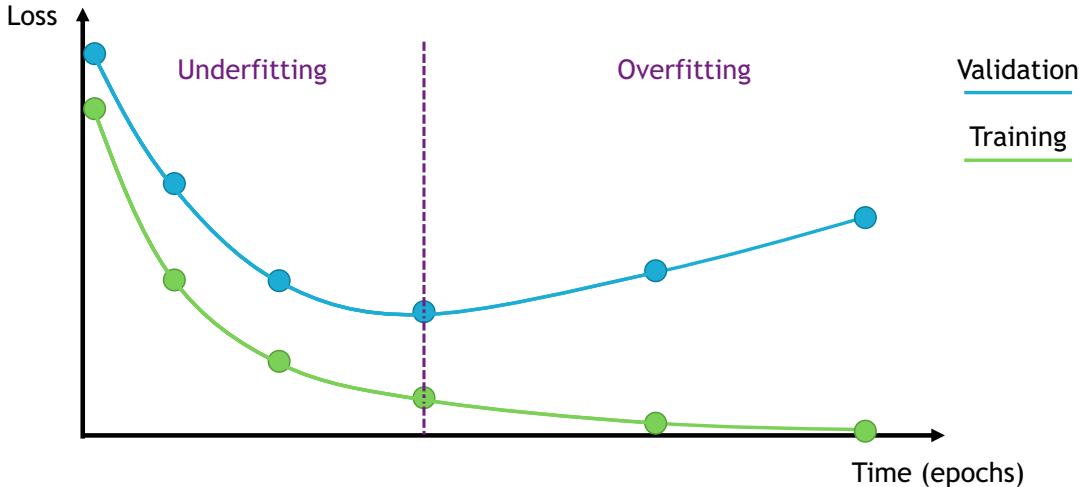


Fig. 8 Illustration of the concept of early stopping. The model that should be selected corresponds to the dashed bar which is the point where the validation loss starts increasing. Before this point, the model is underfitting. After, it is overfitting

neural networks, different optimizers have integrated them into their optimization process in the form of weight decay.

Weight Initialization The way that the weights of neural networks will be initialized is very important, and it can determine whether the algorithm converges at all, with some initial points being so unstable that the algorithm encounters numerical difficulties and fails altogether [24]. Most of the time, the weights are initialized randomly from a Gaussian or uniform distribution. According to [24], the choice of Gaussian or uniform distribution does not seem to matter very much; however, the scale does have a large effect both on the outcome of the optimization procedure and on the ability of the network to generalize. Nevertheless, more tailored approaches have been developed over the last decade that have become the standard initialization points. One of them is the Xavier Initialization [38] which balances between all the layers to have the same activation variance and the same gradient variance. More formally the weights are initialized as:

$$W_{i,j} \sim \text{Uniform}\left(-\sqrt{\frac{6}{m+n}}, \sqrt{\frac{6}{m+n}}\right), \quad (20)$$

where m is the number of inputs and n the number of outputs of matrix W . Moreover, the biases b are initialized to 0.

Drop-out There are other techniques to prevent overfitting, such as drop-out [39], which involves randomly destroying neurons during the training process, thereby reducing the complexity of



Fig. 9 Examples of data transformations applied in the MNIST dataset. Each of these generated samples is considered additional training data

the model. Drop-out is an ensemble method that does not need to build the models explicitly. In practice, at each optimization iteration, random binary masks on the units are considered. The probability of removing a unit (p) is defined as a hyperparameter during the training of the network. During inference, all the units are activated; however, the obtained parameters W are multiplied with this probability p . Drop-out is quite efficient and commonly used in a variety of neural network architectures.

Data Augmentation Since neural networks are data-driven methods, their performance depends on the training data. To increase the amount of data during the training, data augmentation can be performed. It generates slightly modified copies of the existing training data to enrich the training samples. This technique acts as a regularizer and helps reduce overfitting. Some of the most commonly used transformations applied during data augmentation include random rotations, translations, cropping, color jittering, resizing, Gaussian blurring, and many more. In Fig. 9, examples of different transformations on different digits (first column) of the MNIST dataset [40] are presented. For medical images, the TorchIO library allows to easily perform data augmentation [41].

Batch Normalization To ensure that the training of the networks will be more stable and faster, batch normalization has been proposed [42]. In practice, batch normalization re-centers and re-scales the layer's input, mitigating the problem of internal

covariate shift which changes the distribution of the inputs of each layer affecting the learning rate of the network. Even if the method is quite popular, its necessity and use for the training have recently been questioned [43].

3.4 State-of-the-Art Optimizers

Over the years, different optimizers have been proposed and widely used, aiming to provide improvements over the classical stochastic gradient descent. These algorithms are motivated by challenges that need to be addressed with stochastic gradient descent and are focusing on the choice of the proper learning rate, its dynamic change during training, as well as the fact that it is the same for all the parameter updates [44]. Moreover, a proper choice of optimizer could speed up the convergence to the optimal solution. In this subsection, we will discuss some of the most commonly used optimizers nowadays.

3.4.1 Stochastic Gradient Descent with Momentum

One of the limitations of the stochastic gradient descent is that since the direction of the gradient that we are taking is random, it can heavily oscillate, making the training slower and even getting stuck in a saddle point. To deal with this problem, stochastic gradient descent with momentum [45, 46] keeps a history of the previous gradients, and it updates the weights taking into account the previous updates. More formally:

$$\begin{aligned} \mathbf{g}^t &\leftarrow \rho \mathbf{g}^{t-1} + (1 - \rho) G(\mathbf{W}^t) \\ \Delta \mathbf{W}^t &\leftarrow -\eta_t \mathbf{g}^t \\ \mathbf{W}^{t+1} &\leftarrow \mathbf{W}^t + \Delta \mathbf{W}^t \end{aligned}, \quad (21)$$

where \mathbf{g}^t is the direction of the update of the weights in time-step t and $\rho \in [0, 1]$ is a hyperparameter that controls the contribution of the previous gradients and current gradient in the current update. When $\rho = 0$, it is the same as the classical stochastic gradient descent. A large value of ρ will mean that the update is strongly influenced by the previous updates.

The momentum algorithm accumulates an exponentially decaying moving average of the past gradients and continues to move in their direction [24]. Momentum increases the speed of convergence, while it is also helpful to not get stuck in places where the search space is flat (saddle points with zero gradient), since the momentum will pursue the search in the same direction as before the flat region.

3.4.2 AdaGrad

To facilitate and speed up, even more, the training process, optimizers with adaptive learning rates per parameter have been proposed. The adaptive gradient (AdaGrad) optimizer [47] is one of them. It updates each individual parameter proportionally to their component (and momentum) in the gradient. More formally:

$$\begin{aligned}
g^t &\leftarrow G(\mathbf{W}^t) \\
r^t &\leftarrow r^{t-1} + g^t \odot g^t \\
\Delta \mathbf{W}^t &\leftarrow -\frac{\eta}{\delta + \sqrt{r^t}} \odot g^t, \\
\mathbf{W}^{t+1} &\leftarrow \mathbf{W}^t + \Delta \mathbf{W}^t
\end{aligned} \tag{22}$$

where g^t is the gradient estimate vector in time-step t , r^t is the term controlling the per parameter update, and δ is some small quantity that is used to avoid the division by zero. Note that r^t constitutes of the gradient's element-wise product with itself and of the previous term r^{t-1} accumulating the gradients of the previous terms.

This algorithm performs very well for sparse data since it decreases the learning rate faster for the parameters that are more frequent and slower for the infrequent parameters. However, since the update accumulates gradients of the previous steps, the updates could decrease very fast, blocking the learning process. This limitation is mitigated by extensions of the AdaGrad algorithm as we discuss in the next sections.

3.4.3 RMSProp

Another algorithm with adaptive learning rates per parameter is the root mean squared propagation (RMSProp) algorithm, proposed by Geoffrey Hinton. Despite its popularity and use, this algorithm has not been published. RMSProp is an extension of the AdaGrad algorithm dealing with the problem of radically diminishing learning rates by being less influenced by the first iterations of the algorithm. More formally:

$$\begin{aligned}
g^t &\leftarrow G(\mathbf{W}^t) \\
r^t &\leftarrow \rho r^{t-1} + (1 - \rho) g^t \odot g^t \\
\Delta \mathbf{W}^t &\leftarrow -\frac{\eta}{\delta + \sqrt{r^t}} \odot g^t, \\
\mathbf{W}^{t+1} &\leftarrow \mathbf{W}^t + \Delta \mathbf{W}^t
\end{aligned} \tag{23}$$

where ρ is a hyperparameter that controls the contribution of the previous gradients and the current gradient in the current update. Note that RMSProp estimates the squared gradients in the same way as AdaGrad, but instead of letting that estimate continually accumulate over training, we keep a moving average of it, integrating the momentum. Empirically, RMSProp has been shown to be an effective and practical optimization algorithm for deep neural networks [24].

3.4.4 Adam

The effectiveness and advantages of the AdaGrad and RMSProp algorithms are combined in the adaptive moment estimation (Adam) optimizer [48]. The method computes individual adaptive learning rates for different parameters from estimates of the first and second moments of the gradients. More formally:

$$\begin{aligned}
g^t &\leftarrow G(W^t) \\
s^t &\leftarrow \rho_1 s^{t-1} + (1 - \rho_1) g^t \\
r^t &\leftarrow \rho_2 r^{t-1} + (1 - \rho_2) g^t \odot g^t \\
\hat{s}^t &\leftarrow \frac{s^t}{1 - (\rho_1)^t} \\
\hat{r}^t &\leftarrow \frac{r^t}{1 - (\rho_2)^t} \\
\Delta W^t &\leftarrow -\frac{\lambda}{\delta + \sqrt{\hat{r}^t}} \odot \hat{s}^t \\
W^{t+1} &\leftarrow W^t + \Delta W^t
\end{aligned}, \tag{24}$$

where s^t is the gradient with momentum, r^t accumulates the squared gradients with momentum as in RMSProp, and \hat{s}^t and \hat{r}^t are smaller than s^t and r^t , respectively, but they converge toward them. Moreover, δ is some small quantity that is used to avoid the division by zero, while ρ_1 and ρ_2 are hyperparameters of the algorithm. The parameters ρ_1 and ρ_2 control the decay rates of each moving average, respectively, and their value is close to 1. Empirical results demonstrate that Adam works well in practice and compares favorably to other stochastic optimization methods, making it the go-to optimizer for deep learning problems.

3.4.5 Other Optimizers

The development of efficient (in terms of speed and stability) optimizers is still an active research direction. RAdam [49] is a variant of Adam, introducing a term to rectify the variance of the adaptive learning rate. In particular, RAdam leverages a dynamic rectifier to adjust the adaptive momentum of Adam based on the variance and effectively provides an automated warm-up customized to the current dataset to ensure a solid start to training. Moreover, LookAhead [50] was inspired by recent advances in the understanding of loss surfaces of deep neural networks and provides a breakthrough in robust and stable exploration during the entirety of the training. Intuitively, the algorithm chooses a search direction by looking ahead at the sequence of fast weights generated by another optimizer. These are only some of the optimizers that exist in the literature, and depending on the problem and the application, different optimizers could be selected and applied.

4 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a specific category of deep neural networks that employ the convolution operation in order to process the input data. Even though the main concept dates back to the 1990s and is greatly inspired by neuroscience [51] (in particular by the organization of the visual cortex), their widespread use is due to a relatively recent success on the ImageNet Large Scale Visual Recognition Challenge of 2012 [27]. In contrast

to the deep fully connected networks that have been already discussed, CNNs excel in processing data with a spatial or grid-like organization (e.g., time series, images, videos, etc.) while at the same time decreasing the number of trainable parameters due to their weight sharing properties. The rest of this section is first introducing the convolution operation and the motivation behind using it as a building block/module of neural networks. Then, a number of different variations are presented together with examples of the most important CNN architectures. Lastly, the importance of the receptive field – a central property of such networks – will be discussed.

4.1 The Convolution Operation

The convolution operation is defined as the integral of the product of the two functions (f, g)⁵ after one is reversed and shifted over the other function. Formally, we write:

$$h(t) = \int_{-\infty}^{\infty} f(t - \tau)g(\tau) d\tau. \quad (25)$$

Such an operation can also be denoted with an asterisk (*), so it is written as:

$$h(t) = (f * g)(t). \quad (26)$$

In essence, the convolution operation shows how one function affects the other. This intuition arises from the signal processing domain, where it is typically important to know how a signal will be affected by a filter. For example, consider a uni-dimensional continuous signal, like the brain activity of a patient on some electroencephalography electrode, and a Gaussian filter. The result of the convolution operation between these two functions will output the effect of a Gaussian filter on this signal which will, in fact, be a smoothed version of the input.

A different way to think of the convolution operation is that it shows how the two functions are related. In other words, it shows how similar or dissimilar the two functions are at different relative positions. In fact, the convolution operation is very similar to the cross-correlation operation, with the subtle difference being that in the convolution operation, one of the two functions is inverted. In the context of deep learning specifically, the exact differences between the two operations can be of secondary concern; however, the convolution operation has more properties than correlation, such as commutativity. Note also that when the signals are symmetric, both operations will yield the same result.

In order to deal with discrete and finite signals, we can expand the definition of the convolution operation. Specifically, given two

⁵ Note that f and g have no relationship to their previous definitions in the chapter. In particular, f is not the deep learning model.

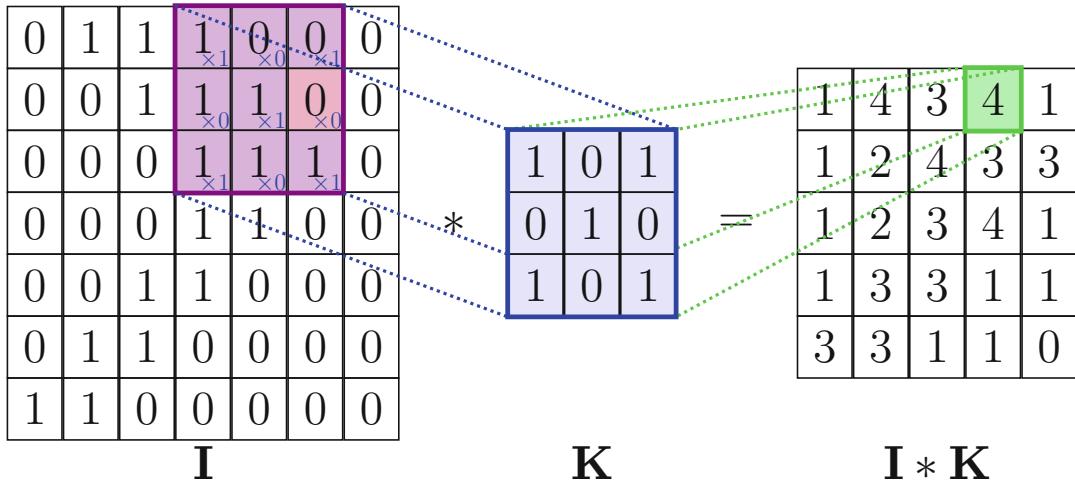


Fig. 10 A visualization of the discrete convolution operation in 2D

discrete signals $f[k]$ and $g[k]$, with $k \in \mathbb{Z}$, the convolution operation is defined by:

$$h[k] = \sum_n f[k-n]g[n]. \quad (27)$$

Lastly, the convolution operation can be extended for multidimensional signals similarly. For example, we can write the convolution operation between two discrete and finite two-dimensional signals (e.g., $I[i, j]$, $K[m, n]$) as:

$$H[i, j] = \sum_m \sum_n I[i-m, j-n]K[m, n]. \quad (28)$$

Very often, the first signal will be the input of interest (e.g., a large size image), while the second signal will be of relatively small size (e.g., a 3×3 or 4×4 matrix) and will implement a specific operation. The second signal is then called a kernel. In Fig. 10, a visualization of the convolution operation is shown in the case of a 2D discrete signal such as an image and a 3×3 kernel. In detail, the convolution kernel is shifted over all locations of the input, and an element-wise multiplication and a summation are utilized to calculate the convolution output at the corresponding location. Examples of applications of convolutions to an image are provided in Fig. 11. Finally, note that, as in multilayer perceptrons, a convolution will generally be followed by a non-linear activation function, for instance, a ReLU (see Fig. 12 for an example of activation applied to a feature map).

In the following sections of this chapter, any reference to the convolution operation will mostly refer to the 2D discrete case. The

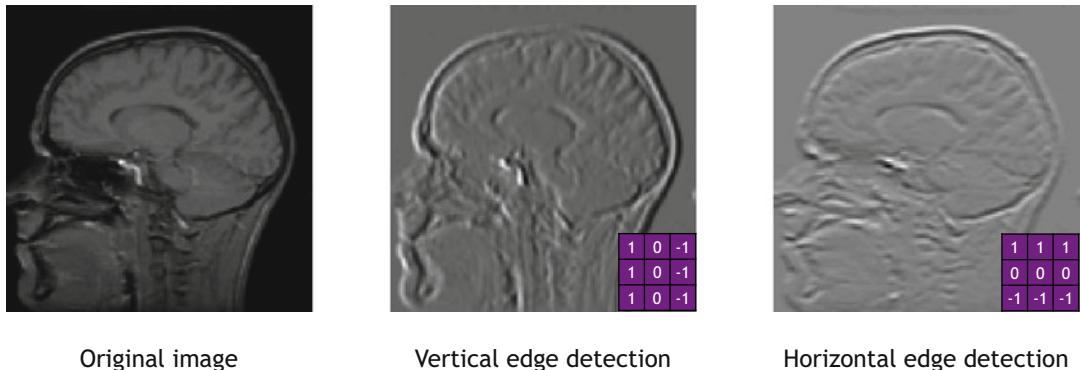


Fig. 11 Two examples of convolutions applied to an image. One of the filters acts as a vertical edge detector and the other one as a horizontal edge detector. Of course, in CNNs, the filters are learned, not predefined, so there is no guarantee that, among the learned filters, there will be a vertical/horizontal case detector, although it will often be the case in practice, especially for the first layers of the architecture

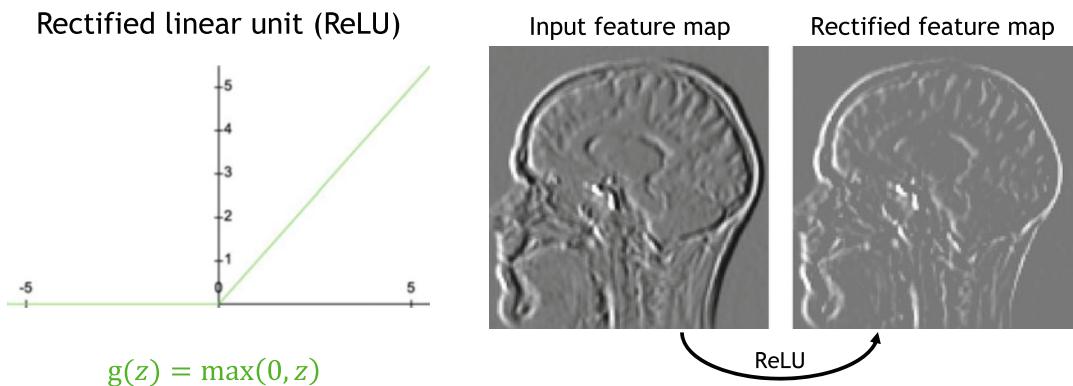


Fig. 12 Example of application of a non-linear activation function (here a ReLU) to an image

extension to the 3D case, which is often encountered in medical imaging, is straightforward.

4.2 Properties of the Convolution Operation

In the case of a discrete domain, the convolution operation can be performed using a simple matrix multiplication without the need of shifting one signal over the other one. This can be essentially achieved by utilizing the Toeplitz matrix transformation. The Toeplitz transformation creates a sparse matrix with repeated elements which, when multiplied with the input signal, produces the convolution result. To illustrate how the convolution operation can be implemented as a matrix multiplication, let's take the example of a 3×3 kernel (K) and a 4×4 input (I):

$$K = \begin{bmatrix} k_{00} & k_{01} & k_{02} \\ k_{10} & k_{11} & k_{12} \\ k_{20} & k_{21} & k_{22} \end{bmatrix} \quad \text{and} \quad I = \begin{bmatrix} i_{00} & i_{01} & i_{02} & i_{03} \\ i_{10} & i_{11} & i_{12} & i_{13} \\ i_{20} & i_{21} & i_{22} & i_{23} \\ i_{30} & i_{31} & i_{32} & i_{33} \end{bmatrix}.$$

Then, the convolution operation can be computed as a matrix multiplication between the Toeplitz transformed kernel:

$$K = \begin{bmatrix} k_{00} & k_{01} & k_{02} & 0 & k_{10} & k_{11} & k_{12} & 0 & k_{20} & k_{21} & k_{22} & 0 & 0 & 0 & 0 & 0 \\ 0 & k_{00} & k_{01} & k_{02} & 0 & k_{10} & k_{11} & k_{12} & 0 & k_{20} & k_{21} & k_{22} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & k_{00} & k_{01} & k_{02} & 0 & k_{10} & k_{11} & k_{12} & 0 & k_{20} & k_{21} & k_{22} & 0 \\ 0 & 0 & 0 & 0 & 0 & k_{00} & k_{01} & k_{02} & 0 & k_{10} & k_{11} & k_{12} & 0 & k_{20} & k_{21} & k_{22} \end{bmatrix}$$

and a reshaped input:

$$I = [i_{00} \ i_{01} \ i_{02} \ i_{03} \ i_{10} \ i_{11} \ i_{12} \ i_{13} \ i_{20} \ i_{21} \ i_{22} \ i_{23} \ i_{30} \ i_{31} \ i_{32} \ i_{33}]^\top.$$

The produced output will need to be reshaped as a 2×2 matrix in order to retrieve the convolution output. This matrix multiplication implementation is quite illuminating on a few of the most important properties of the convolution operation. These properties are the main motivation behind using such elements in deep neural networks.

By transforming the convolution operation to a matrix multiplication operation, it is evident that it can fit in the formalization of the linear functions, which has already been presented in Subheading 2.3. As such, deep neural networks can be designed in a way to utilize trainable convolution kernels. In practice, multiple convolution kernels are learned at each convolutional block, while several of these trainable convolutional blocks are stacked on top of each other forming deep CNNs. Typically, the output of a convolution operation is called a *feature map* or just *features*.

Another important aspect of the convolution operation is that it requires much fewer parameters than the fully connected MLP-based deep neural networks. As it can also be seen from the K matrix, the exact same parameters are shared across all locations. Eventually, rather than learning a different set of parameters for the different locations of the input, only one set is learned. This is referred to as *parameter sharing* or *weight sharing* and can greatly decrease the amount of memory that is required to store the network parameters. An illustration of the process of *weight sharing* across locations, together with the fact that multiple filters (resulting in multiple feature maps) are computed for a given layer, is illustrated in Fig. 13. The multiple feature maps for a given layer are stored using another dimension (see Fig. 14), thus resulting in a 3D

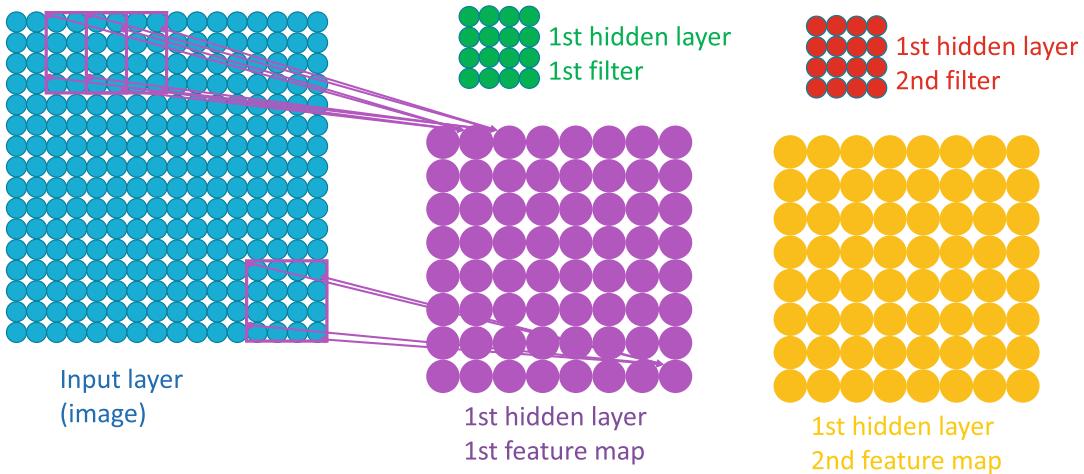


Fig. 13 For a given layer, several (usually many) filters are learned, each of them being able to detect a specific characteristic in the image, resulting in several feature/filter maps. On the other hand, for a given filter, the weights are shared across all the locations of the image

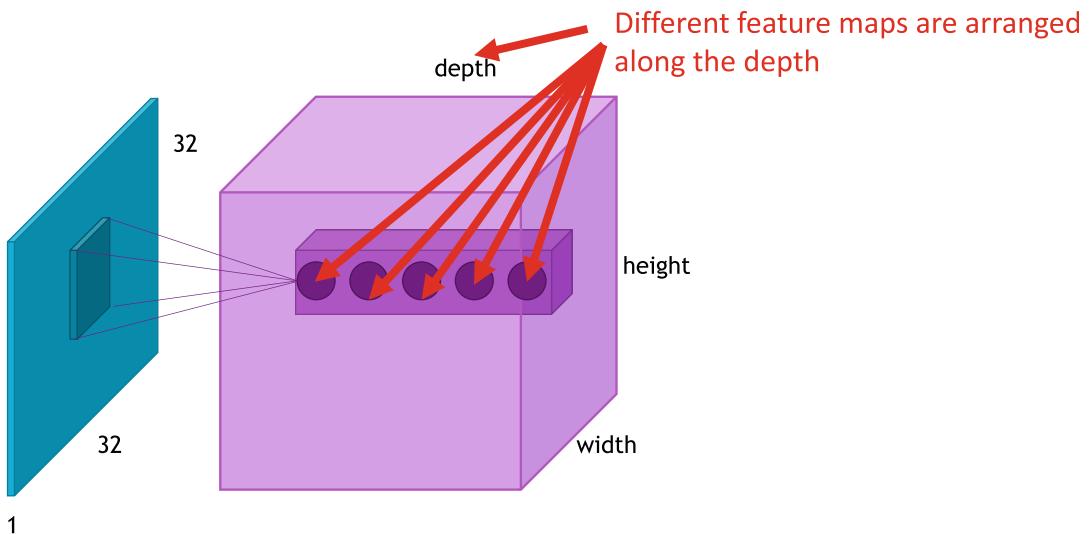


Fig. 14 The different feature maps for a given layer are arranged along another dimension. The feature maps will thus be a 3D array when the input is a 2D image (and a 4D array when the input is a 3D image)

array when the input is a 2D image (and a 4D array when the input is a 3D image).

Convolutional neural networks have proven quite powerful in processing data with spatial structure (e.g., images, videos, etc.). This is effectively based on the fact that there is a local connectivity of the kernel elements while at the same time the same kernel is applied at different locations of the input. Such processing grants a quite useful property called *translation equivariance* enabling the

network to output similar responses at different locations of the input. An example of the usefulness of such a property can be identified on an image detection task. Specifically, when training a network to detect tumors in an MR image of the brain, the model should respond similarly regardless of the location where the anomaly can be manifested.

Lastly, another important property of the convolution operation is that it decouples the size of the input with the trainable parameters. For example, in the case of MLPs, the size of the weight matrix is a function of the dimension of the input. Specifically, a densely connected layer that maps 256 features to 10 outputs would have a size of $W \in \mathbb{R}^{10 \times 256}$. On the contrary, in convolutional layers, the number of trainable parameters only depends on the kernel size and the number of kernels that a layer has. This eventually allows the processing of arbitrarily sized inputs, for example, in the case of fully convolutional networks.

4.3 Functions and Variants

An observant reader might have noticed that the convolution operation can change the dimensionality of the produced output. In the example visualized in Fig. 10, the image of size 7×7 , when convolved with a kernel of size 3×3 , produces a feature map of size of 5×5 . Even though dimension changes can be avoided with appropriate padding (see Fig. 15 for an illustration of this process) prior to the convolution operation, in some cases, it is actually desired to reduce the dimensions of the input. Such a decrease can be achieved in a number of ways depending on the task at hand. In this subsection, some of the most typical functions that are utilized in CNNs will be discussed.

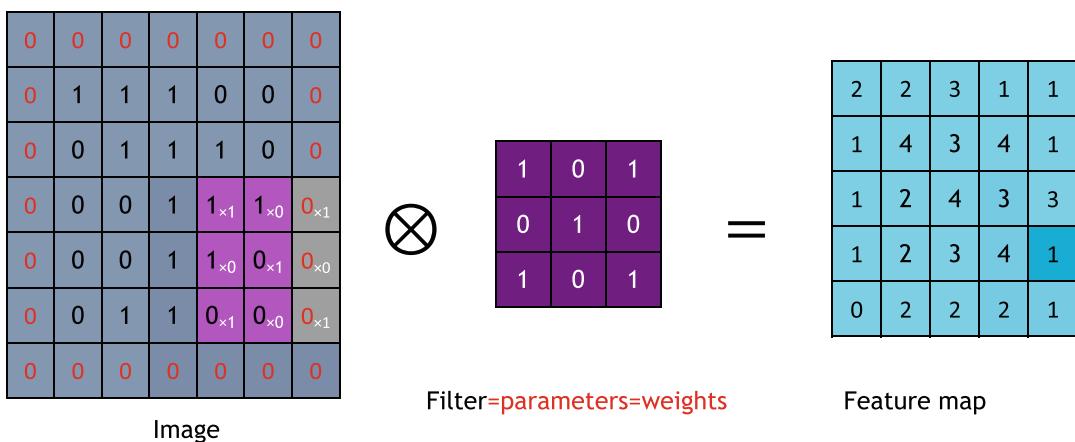


Fig. 15 The padding operation, which involves adding zeros around the image, allows to obtain feature maps that are of the same size as the original image

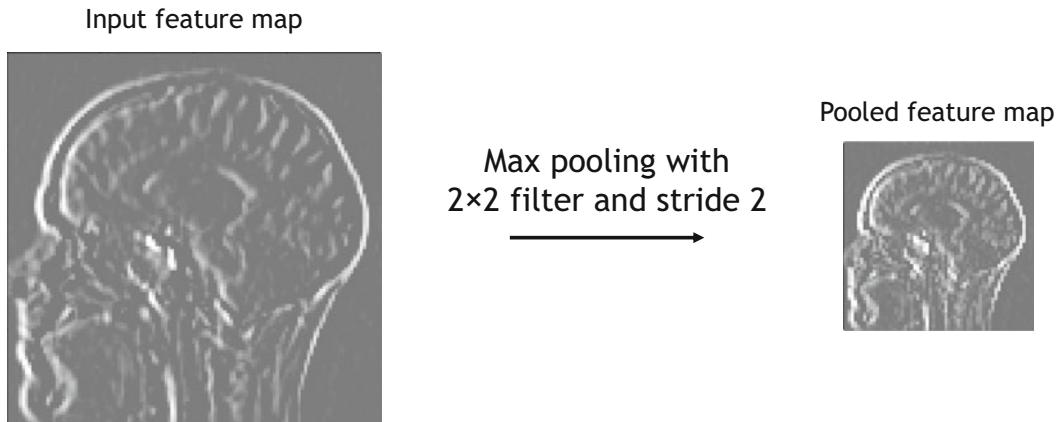


Fig. 16 Effect of a pooling operation. Here, a maximum pooling of size 2×2 with a stride of 2

Downsampling Operations (i.e., Pooling Layers) In many CNN architectures, there is an extensive use of downsampling operations that aim to compress the size of the feature maps and decrease the computational burden. Otherwise referred to as pooling layers, these processing operations are aggregating the values of their input depending on their design. Some of the most common downsampling layers are the *maximum pooling*, *average pooling*, or *global average pooling*. In the first two, either the maximum or the average value is used as a feature for the output across non-overlapping regions of a predefined pooling size. In the case of the global average pooling, the spatial dimensions are all represented with the average value. An example of pooling is provided in Fig. 16.

Strided Convolution The strided convolution refers to the specific case in which, instead of applying the convolution operation for every location using a step size (or stride, s) of 1, different step sizes can be considered (Fig. 17). Such an operation will produce a convolution output with much fewer elements. Convolutional blocks with $s > 1$ can be found on CNN architectures as a way to decrease the feature sizes in intermediate layers.

Atrous or Dilated Convolution Dilated, also called atrous, convolution is the convolution with kernels that have been dilated by inserting zero holes (*à trous* in French) between the non-zero values of a kernel. In this case, an additional parameter (d) of the convolution operation is added, and it is changing the distance between the kernel elements. In essence, it is increasing the reach of the kernel but keeping the number of trainable parameters the same. For example, a dilated convolution with a kernel size of 3×3 and a dilation rate of $d=2$ would be sparsely arranged on a 5×5 grid.

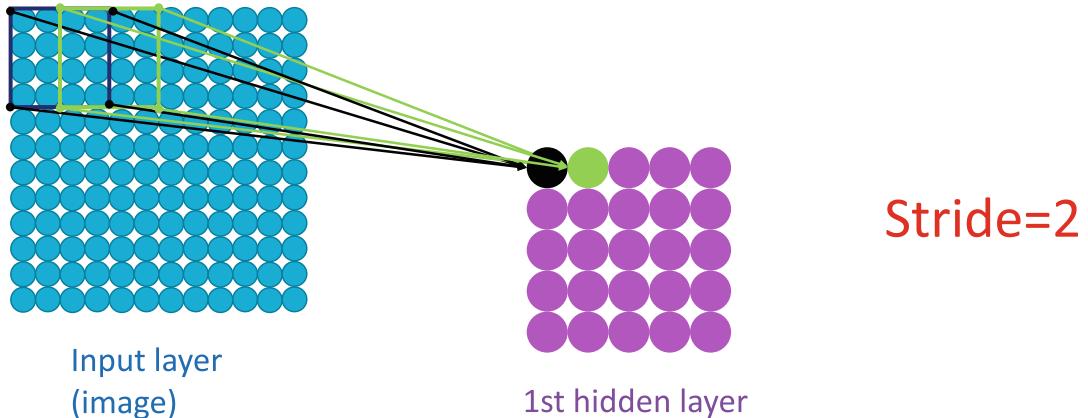


Fig. 17 Stride operation, here with a stride of 2

Transposed Convolution In certain circumstances, one needs not only to downsample the spatial dimensions of the input but also, usually at a later stage of the network, apply an upsample operation. The most emblematic case is the task of image segmentation (*see* Chap. 13), in which a pixel-level classification is expected, and therefore, the output of the neural network should have the same size as the input. In such cases, several upsampling operations are typically applied. The upsampling can be achieved by a transposed convolution operation that will eventually increase the size of the output. In details, the transposed convolution is performed by dilating the input instead of the kernel before applying a convolution operation. In this way, an input of size 5×5 will reach a size of 10×10 after being dilated with $d = 2$. With proper padding and using a kernel of size 3×3 , the output will eventually double in size.

4.4 Receptive Field Calculation

In the context of deep neural networks and specifically CNNs, the term receptive field is used to define the proportion of the input that produces a specific feature. For example, a CNN that takes an image as input and applies only a single convolution operation with a kernel size of 3×3 would have a receptive field of 3×3 . This means that for each pixel of the first feature map, a 3×3 region of the input would be considered. Now, if another layer were to be added, with again 3×3 size, then the receptive field of the new feature map with respect to the CNN's input would be 5×5 . In other words, the proportion of the input that is used to calculate each element of the feature map of the second convolution layer increases.

Calculating the receptive field at different parts of a CNN is crucial when trying to understand the inner workings of a specific architecture. For instance, a CNN that is designed to take as an input an image of size 256×256 and that requires information

from all parts of it should have a receptive field close to the size of the input. The receptive field can be influenced by all the different convolution parameters and down-/upsampling operations described in the previous section. A comprehensive presentation of mathematical derivations for calculating receptive fields for CNNs is given in [52].

4.5 Classical Convolutional Neural Network Architectures

In the last decades, a variety of convolutional neural network architectures have been proposed. In this chapter, we cover only a few classical architectures for classification and regression. Note that classification and regression can usually be performed with the same architecture, just changing the loss function (e.g., cross-entropy for classification, mean squared error for regression). Architectures for other tasks can be found in other chapters.

A Basic CNN Architecture Let us start with the most simple CNN, which is actually very close to the original one proposed by LeCun et al. [53], sometimes called “LeNet.” Such architecture is typically composed of two parts: the first one is based on convolution operations and learns the features for the image and the second part flattens the features and inputs them to a set of fully connected layers (in other words, a multilayer perceptron) for performing the classification/regression (*see* illustration in Fig. 18). Note that, of course, the whole network is trained end to end: the two parts are not trained independently. In the first part, one combines a series of blocks composed of a convolution operation (possibly strided and/or dilated), a non-linear activation function (for instance, a ReLU), and a pooling operation. It is often a good idea to include a drawing of the different layers of the chosen architecture.

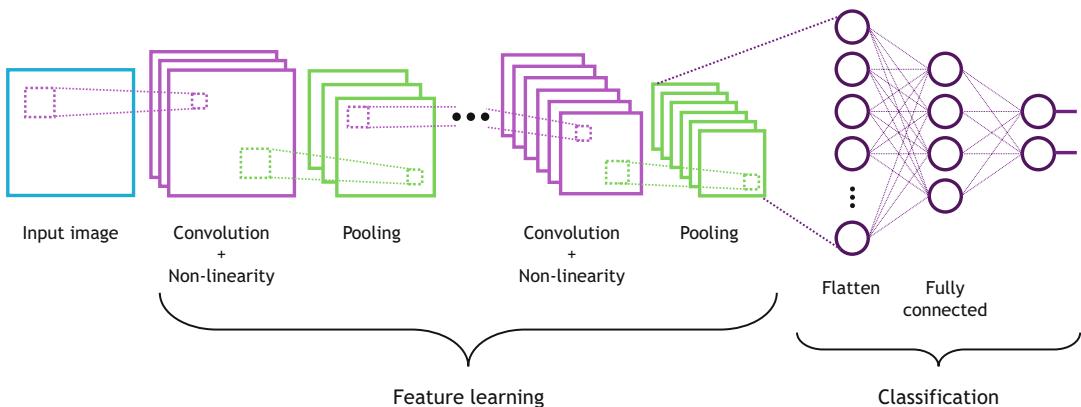


Fig. 18 A basic CNN architecture. Classically, it is composed of two main parts. The first one, using convolution operations, performs feature learning. The features are then flattened and fed into a set of fully connected layers (i.e., a multilayer perceptron), which performs the classification or the regression task

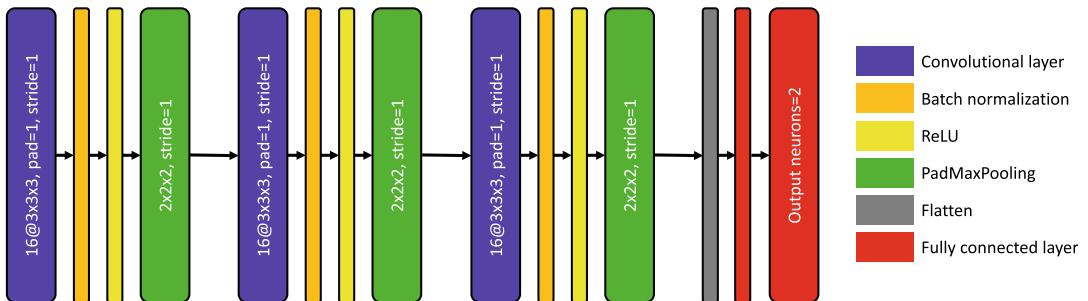


Fig. 19 A drawing describing a CNN architecture. Classically, it is composed of two main parts. Here $16@3 \times 3 \times 3$ means that 16 features with a $3 \times 3 \times 3$ convolution kernel will be computed. For the pooling operation, the kernel size is also mentioned (2×2). Finally, the stride is systematically indicated

Unfortunately, there is no harmonized format for such a description. An example is provided in Fig. 19.

One of the first CNN architectures that follow this paradigm is the AlexNet architecture [54]. AlexNet was one of the first papers that empirically indicated that the ReLU activation function makes the convergence of CNNs faster compared to other non-linearities such as the tanh. Moreover, it was the first architecture that achieved a top 5 error rate of 18.2% on the ImageNet dataset, outperforming all the other methods on this benchmark by a huge margin (about 10%). Prior to AlexNet, best-performing methods were using (very sophisticated) pre-extracted features and classical machine learning. After this advance, deep learning in general and CNNs, in particular, became very active research directions to address different computer vision problems. This resulted in the introduction of a variety of architectures such as VGG16 [55] that reported a 7.3% error rate on ImageNet, introducing some changes such as the use of smaller kernel filters. Following these advances, and even if there were a lot of different architectures proposed during that period, one could mention the Inception architecture [56], which was one of the deepest architectures of that period and which further reduced the error rate on ImageNet to 6.7%. One of the main characteristics of this architecture was the inception modules, which applied multiple kernel filters of different sizes at each level of the architecture. To solve the problem of vanishing gradients, the authors introduced auxiliary classifiers connected to intermediate layers, expecting to encourage discrimination in the lower stages in the classifier, increasing the gradient signal that gets propagated back, and providing additional regularization. During inference, these classifiers were completely discarded.

In the following section, some other recent and commonly used CNN architectures, especially for medical applications, will be presented.

ResNet One of the most commonly used CNN architectures, even today, is the ResNet [57]. ResNet reduced the error rate on ImageNet to 3.6%, while it was the first deep architecture that proposed novel concepts on how to gracefully go deeper than a few dozen of layers. In particular, the authors introduced a deep residual learning framework. The main idea of this residual learning is that instead of learning the desired underlying mapping of each network level, they learn the residual mapping. More formally, instead of learning the $H(x)$ mapping after the convolutional and non-linear layers, they fit another mapping of $F(x) = H(x) - x$ on which the original mapping is recast into $F(x) + x$. Feedforward neural networks can realize this mapping with “shortcut connections” by simply performing identity mapping, and their outputs are added to the outputs of the stacked layers. Such identity connections add neither additional complexity nor parameters to the network, making such architectures extremely powerful.

Different ResNet architectures have been proposed even in the original paper. Even though the depth of the network is increased with the additional convolutions, especially for the 152-layer ResNet (11.3 billion floating point operations), it still has lower complexity (i.e., fewer parameters) than VGG16/VGG19 networks. Currently, different layered-size ResNet architectures pre-trained on ImageNet are used as backbones for various problems and applications, including medical imaging. Pre-trained ResNet models, even if they are 2D architectures, are commonly used on histopathology [58, 59], chest X-ray [60], or even brain imaging [61, 62], while the way that such pre-trained networks work for medical applications gathered the attention of different studies such as [63]. However, it should be noted that networks pre-trained on ImageNet are not always efficient for medical imaging tasks, and there are cases where they perform poorly, much lower than simpler CNNs trained from scratch [64]. Nevertheless, a pre-trained ResNet is very often a good idea to use for a first try in a given application. Finally, there was an effort from the medical community to train 3D variations of ResNet architectures on a large amount of 3D medical data and release the pre-trained models. Such an effort is presented in [65] in which the authors trained and released different 3D ResNet architectures trained on different publicly available 3D datasets, including different anatomies such as the brain, prostate, liver, heart, and pancreas.

EfficientNet A more recent CNN architecture that is worth mentioning in this section is the recently presented EfficientNet [66]. EfficientNets are a family of neural networks that are balancing all dimensions of the network (width/depth/resolution) automatically. In particular, the authors propose a simple yet effective compound scaling method for obtaining these hyperparameters. In particular, the main compound coefficient ϕ uniformly scales

network width, depth, and resolution in a principled way: depth = α^ϕ , width = β^ϕ , resolution = γ^ϕ s.t. $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$, $\alpha \geq 1$, $\beta \geq 1$, $\gamma \geq 1$. In this formulation, the parameters α, β, γ are constants, and a small grid search can determine them. This grid search resulted in *eight* different architectures presented in the original paper. EfficientNet is used more and more for medical imaging tasks, as can be seen in multiple recent studies [67–69].

5 Autoencoders

An autoencoder is a type of neural network that can learn a compressed representation (called the latent space representation) of the training data. As opposed to the multilayer perceptrons and CNNs seen until now that are used for supervised learning, auto-encoders have widely been used for unsupervised learning, with a wide range of applications. The architecture of autoencoders is composed of a contracting path (called the encoder), which will transform the input into a lower-dimensional representation, and an expanding path (called the decoder), which will aim at reconstructing the input as well as possible from the lower-dimensional representation (see Fig. 20).

The loss is usually the ℓ_2 loss and the cost function is then:

$$J(\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{i=1}^n \| \mathbf{x}^{(i)} - D_{\boldsymbol{\theta}}(E_{\boldsymbol{\phi}}(\mathbf{x}^{(i)})) \|_2^2, \quad (29)$$

where $E_{\boldsymbol{\phi}}$ is the encoder (and $\boldsymbol{\phi}$ its parameters) and $D_{\boldsymbol{\theta}}$ is the decoder (and $\boldsymbol{\theta}$ its parameters). Note that, in Fig. 20, $D_{\boldsymbol{\theta}}(E_{\boldsymbol{\phi}}(\mathbf{x}))$ is denoted as $\hat{\mathbf{x}}$. More generally, one can write:

$$J(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{\mathbf{x} \sim \mu_{\text{ref}}} [d(\mathbf{x}, D_{\boldsymbol{\theta}}(E_{\boldsymbol{\phi}}(\mathbf{x})))], \quad (30)$$

where μ_{ref} is the reference distribution that one is trying to approximate and d is the reconstruction function. When μ_{ref} is the

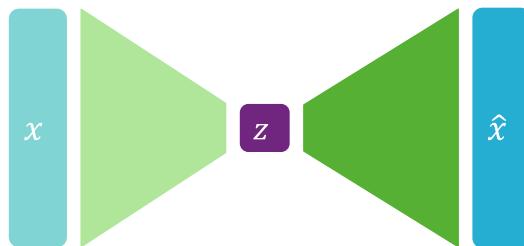


Fig. 20 The general principle of a denoising autoencoder. It aims at learning of a low-dimensional representation (latent space) \mathbf{z} of the training data. The learning is done by aiming to provide a faithful reconstruction $\hat{\mathbf{x}}$ of the input data \mathbf{x}

empirical distribution of the training set and d is the ℓ_2 norm, Eq. 30 is equivalent to Eq. 29.

Many variations of autoencoders exist, to prevent autoencoders from learning the identity function and to improve their ability to capture important information and learn richer representations. Among them, *sparse autoencoders* offer an alternative method for introducing an information bottleneck without requiring a reduction in the number of nodes at the hidden features. This is done by constructing the loss function such that it penalizes activations within a layer. This is achieved by enforcing sparsity in the network and encouraging it to learn an encoding and decoding which relies only on activating a small number of neurons. This sparsity is enforced in two main ways, an ℓ_1 regularization on the parameters of the network and a Kullback-Leibler divergence, which is a measure of the difference between two probability distributions. More information about sparse autoencoders could be found in [70]. Moreover, a quite common type of autoencoders is the *denoising autoencoders* [71], on which the model is tasked with reproducing the input as closely as possible while passing through some sort of information bottleneck (Fig. 20). This way, the model is not able to simply develop a mapping that memorizes the training data but rather learns a vector field for mapping the input data toward a lower-dimensional manifold. One should note here that the vector field is typically well-behaved in the regions where the model has observed data during training. In out-of-distribution data, the reconstruction error is both large and does not always point in the direction of the true distribution. This observation makes these networks quite popular for anomaly detection in medical data [72]. Additionally, *contractive autoencoders* [73] are other variants of this type of models, adding the contractive regularization loss to the standard autoencoder loss. Intuitively, it forces very similar inputs to have a similar encoding, and in particular, it requires the derivative of the hidden layer activations to be small with respect to small changes in the input. The denoising autoencoders can be understood as a variation of the contractive autoencoder. In the limit of small Gaussian noise, the denoising autoencoders make the reconstruction error resistant to finite-sized input perturbations, while the contractive autoencoders make the extracted features resistant to small input perturbations.

Depending on the input type, different autoencoder architectures could be designed. In particular, when the inputs are images, the encoder and the decoder are classically composed of convolutional blocks. The decoder uses, for instance, transposed convolutions to perform the expansion. Finally, the addition of skip connections has led to the U-Net [74] architectures that are commonly used for segmentation purposes. Segmentation architectures will be more extensively described in Chap. 13. Finally, variational autoencoders, which rely on a different mathematical formulation,

are not covered in the present chapter and are presented, together with other generative models, in Chap. 5.

6 Conclusion

Deep learning is a very fast evolving field, with numerous still unanswered theoretical questions. However, deep learning-based models have become the state-of-the-art methods for a variety of fields and tasks. In this chapter, we presented the basic principles of deep learning, covering both perceptrons and convolutional neural networks. All architectures were feedforward and recurrent networks are covered in Chap. 4. Generative adversarial networks are covered in Chap. 5, along with other generative models. Chapter 6 presents a recent class of deep learning methods, which does not use convolutions, and that are called transformers. Finally, throughout the other chapters of the book, different deep learning architectures are presented for various types of applications.

Acknowledgements

This work was supported in part by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), reference ANR-10-IAIHU-06 (Institut Hospitalo-Universitaire ICM), and ANR-21-CE45-0007 (Hagnodice).

References

- Rosenblatt F (1957) The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory, Buffalo
- Minsky M, Papert S (1969) Perceptron: an introduction to computational geometry. MIT Press, Cambridge, MA
- Minsky ML, Papert SA (1988) Perceptrons: expanded edition. MIT Press, Cambridge, MA
- Linnainmaa S (1976) Taylor expansion of the accumulated rounding error. BIT Numer Math 16(2):146–160
- Werbos PJ (1982) Applications of advances in nonlinear sensitivity analysis. In: System modeling and optimization. Springer, Berlin, pp 762–770
- Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088): 533–536
- Le Cun Y (1985) Une procédure d’apprentissage pour réseau à seuil assymétrique. Cognitiva 85:599–604
- Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8): 1735–1780
- Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
- Hinton GE (2007) Learning multiple layers of representation. Trends Cogn Sci 11(10): 428–434
- Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 248–255
- Bergstra J, Bastien F, Breuleux O, Lamblin P, Pascanu R, Delalleau O, Desjardins G, Warde-Farley D, Goodfellow I, Bergeron A et al

- (2011) Theano: deep learning on GPUs with Python. In: NIPS 2011, Big learning workshop, Granada, Spain, vol 3. Citeseer, pp 1–48
13. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia, pp 675–678
 14. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M et al (2016) TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467
 15. Chollet F et al (2015) Keras. <https://github.com/fchollet/keras>
 16. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) PyTorch: an imperative style, high-performance deep learning library. In: Advances in neural information processing systems, vol 32
 17. Hebb DO (1949) The organization of behavior: a psychological theory. Wiley, New York
 18. Cybenko G (1989) Approximations by superpositions of a sigmoidal function. Math Control Signals Syst 2:183–192
 19. Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2(5):359–366
 20. Mhaskar HN (1996) Neural networks for optimal approximation of smooth and analytic functions. Neural Comput 8(1):164–177
 21. Pinkus A (1999) Approximation theory of the MLP model in neural networks. Acta Numer 8: 143–195
 22. Poggio T, Mhaskar H, Rosasco L, Miranda B, Liao Q (2017) Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. Int J Autom Comput 14(5):503–519
 23. Rolnick D, Tegmark M (2017) The power of deeper networks for expressing natural functions. arXiv preprint arXiv:170505502
 24. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge, MA
 25. Cover TM (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans Electron Comput 3:326–334
 26. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics, JMLR workshop and conference proceedings, pp 315–323
 27. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, vol 25
 28. Hein M, Andriushchenko M, Bitterwolf J (2019) Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 41–50
 29. Maas AL, Hannun AY, Ng AY et al (2013) Rectifier nonlinearities improve neural network acoustic models. In: Proc. ICML, Atlanta, Georgia, vol 30. p 3
 30. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE international conference on computer vision, pp 1026–1034
 31. Ramachandran P, Zoph B, Le QV (2017) Searching for activation functions. arXiv preprint arXiv:171005941
 32. Dauphin YN, Pascanu R, Gulcehre C, Cho K, Ganguli S, Bengio Y (2014) Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Advances in neural information processing systems, vol 27
 33. Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Springer, Berlin, pp 177–186
 34. Allen-Zhu Z, Li Y, Song Z (2019) A convergence theory for deep learning via over-parameterization. In: International conference on machine learning, PMLR, pp 242–252
 35. Baydin AG, Pearlmutter BA, Radul AA, Siskind JM (2018) Automatic differentiation in machine learning: a survey. J Mach Learn Res 18:1–43
 36. Prechelt L (1998) Early stopping—but when? In: Neural networks: tricks of the trade. Springer, Berlin, pp 55–69
 37. Reed R, MarksII RJ (1999) Neural smithing: supervised learning in feedforward artificial neural networks. MIT Press, Cambridge, MA
 38. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR workshop and conference proceedings, pp 249–256
 39. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from

- overfitting. *J Mach Learn Res* 15(1): 1929–1958
40. Deng L (2012) The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Process Mag* 29(6): 141–142
 41. Pérez-García F, Sparks R, Ourselin S (2021) TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput Methods Programs Biomed* 208: 106236
 42. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, PMLR, pp 448–456
 43. Brock A, De S, Smith SL, Simonyan K (2021) High-performance large-scale image recognition without normalization. In: International conference on machine learning, PMLR, pp 1059–1071
 44. Ruder S (2016) An overview of gradient descent optimization algorithms. arXiv preprint arXiv:160904747
 45. Polyak BT (1964) Some methods of speeding up the convergence of iteration methods. *USSR Comput Math Math Phys* 4(5):1–17
 46. Qian N (1999) On the momentum term in gradient descent learning algorithms. *Neural Netw* 12(1):145–151
 47. Duchi J, Hazan E, Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. *J Mach Learn Res* 12(7)
 48. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980
 49. Liu L, Jiang H, He P, Chen W, Liu X, Gao J, Han J (2019) On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:190803265
 50. Zhang M, Lucas J, Ba J, Hinton GE (2019) LookAhead optimizer: k steps forward, 1 step back. *Adv Neural Inf Process Syst* 32
 51. Fukushima K, Miyake S (1982) Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition. In: Competition and cooperation in neural nets. Springer, Berlin, pp 267–285
 52. Araujo A, Norris W, Sim J (2019) Computing receptive fields of convolutional neural networks. Distill <https://doi.org/10.23915/distill.00021>
 53. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. *Neural Comput* 1(4): 541–551
 54. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges C, Bottou L, Weinberger K (eds) Advances in neural information processing systems, vol 25. Curran Associates. <https://proceedings.neurips.cc/paper/2012/file/c39862d3b9d6b76c8436e924a68c45b-Paper.pdf>
 55. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556
 56. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
 57. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
 58. Lu MY, Williamson DF, Chen TY, Chen RJ, Barbieri M, Mahmood F (2021) Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng* 5(6):555–570
 59. Benkirane H, Vakalopoulou M, Christodoulidis S, Garberis IJ, Michiels S, Cournéde PH (2022) Hyper-AdaC: adaptive clustering-based hypergraph representation of whole slide images for survival analysis. In: Machine learning for health, PMLR, pp 405–418
 60. Horry MJ, Chakraborty S, Paul M, Ulhaq A, Pradhan B, Saha M, Shukla N (2020) X-ray image based COVID-19 detection using pre-trained deep learning models. Engineering Archive, Menomonie
 61. Li JP, Khan S, Alshara MA, Alotaibi RM, Mawuli C et al (2022) DACBT: deep learning approach for classification of brain tumors using MRI data in IoT healthcare environment. *Sci Rep* 12(1):1–14
 62. Nandhini I, Manjula D, Sugumaran V (2022) Multi-class brain disease classification using modified pre-trained convolutional neural networks model with substantial data augmentation. *J Med Imaging Health Inform* 12(2): 168–183
 63. Raghu M, Zhang C, Kleinberg J, Bengio S (2019) Transfusion: understanding transfer learning for medical imaging. In: Advances in neural information processing systems, vol 32

64. Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Dormont D, Durriezman S, Burgos N, Colliot O (2020) Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. *Med Image Anal* 63:101694
65. Chen S, Ma K, Zheng Y (2019) Med3D: transfer learning for 3D medical image analysis. arXiv preprint arXiv:190400625
66. Tan M, Le Q (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning, PMLR, pp 6105–6114
67. Wang J, Liu Q, Xie H, Yang Z, Zhou H (2021) Boosted EfficientNet: detection of lymph node metastases in breast cancer using convolutional neural networks. *Cancers* 13(4):661
68. Oloko-Oba M, Viriri S (2021) Ensemble of EfficientNets for the diagnosis of tuberculosis. *Comput Intell Neurosci* 2021:9790894
69. Ali K, Shaikh ZA, Khan AA, Laghari AA (2021) Multiclass skin cancer classification using EfficientNets—a first step towards preventing skin cancer. *Neurosci Inform* 2(4):100034
70. Ng A et al (2011) Sparse autoencoder. CS294A Lecture Notes 72(2011):1–19
71. Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on machine learning, pp 1096–1103
72. Baur C, Denner S, Wiestler B, Navab N, Albarqouni S (2021) Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. *Med Image Anal* 69: 101952
73. Salah R, Vincent P, Muller X, et al (2011) Contractive auto-encoders: explicit invariance during feature extraction. In: Proceedings of the 28th international conference on machine learning, pp 833–840
74. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, Berlin, pp 234–241

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made. The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 4

Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Introduction to Influential Research

Susmita Das, Amara Tariq, Thiago Santos, Sai Sandeep Kantareddy, and Imon Banerjee

Abstract

Recurrent neural networks (RNNs) are neural network architectures with hidden state and which use feedback loops to process a sequence of data that ultimately informs the final output. Therefore, RNN models can recognize sequential characteristics in the data and help to predict the next likely data point in the data sequence. Leveraging the power of sequential data processing, RNN use cases tend to be connected to either language models or time-series data analysis. However, multiple popular RNN architectures have been introduced in the field, starting from SimpleRNN and LSTM to deep RNN, and applied in different experimental settings. In this chapter, we will present six distinct RNN architectures and will highlight the pros and cons of each model. Afterward, we will discuss real-life tips and tricks for training the RNN models. Finally, we will present four popular language modeling applications of the RNN models –text classification, summarization, machine translation, and image-to-text translation– thereby demonstrating influential research in the field.

Key words Recurrent neural network (RNN), LSTM, GRU, Bidirectional RNN (BRNN), Deep RNN, Language modeling

1 Introduction

Recurrent neural network (RNN) is a specialized neural network with feedback connection for processing sequential data or time-series data in which the output obtained is fed back into it as input along with the new input at every time step. The feedback connection allows the neural network to remember the past data when processing the next output. Such processing can be defined as a recurring process, and hence the architecture is also known as recurring neural network.

RNN concept was first proposed by Rumelhart et al. [1] in a letter published by Nature in 1986 to describe a new learning procedure with a self-organizing neural network. Another important historical moment for RNNs is the (re-)discovery of Hopfield

networks which is a special kind of RNN with symmetric connections where the weight from one node to another and from the latter to the former are the same (symmetric). The Hopfield network [2] is fully connected, so every neuron's output is an input to all the other neurons, and updating of nodes happens in a binary way (0/1). These types of networks were specifically designed to simulate the human memory.

The other types of RNNs are input-output mapping networks, which are used for classification and prediction of sequential data. In 1993, Schmidhuber et al. [3] demonstrated credit assignment across the equivalent of 1,200 layers in an unfolded RNN and revolutionized sequential modeling. In 1997, one of the most popular RNN architectures, the long short-term memory (LSTM) network which can process long sequences, was proposed.

In this chapter, we summarize the six most popular contemporary RNN architectures and their variations and highlight the pros and cons of each. We also discuss real-life tips and tricks for training the RNN models, including various skip connections and gradient clipping. Finally, we highlight four popular language modeling applications of the RNN models –text classification, summarization, machine translation, and image-to-text translation– thereby demonstrating influential research in each area.

2 Popular RNN Architectures

In addition to the SimpleRNN architecture, many variations were proposed to address different use cases. In this section, we will unwrap some of the popular RNN architectures like LSTM, GRU, bidirectional RNN, deep RNN, and attention models and discuss their pros and cons.

2.1 SimpleRNN

SimpleRNN architecture, which is also known as SimpleRNN, contains a simple neural network with a feedback connection. It has the capability to process sequential data of variable length due to the parameter sharing which generalizes the model to process sequences of variable length. Unlike feedforward neural networks which have separate weights for each input feature, RNN shares the same weights across several time steps. In RNN, the output of a present time step depends on the previous time steps and is obtained by the same update rule which is used to obtain the previous outputs. As we will see, the RNN can be unfolded into a deep computational graph in which the weights are shared across time steps.

The RNN operating on an input sequence $\mathbf{x}^{(t)}$ with a time step index t ranging from 1 to τ is illustrated in Fig. 1. The time step index t may not necessarily refer to the passage of time in the real world; it can refer to the position in the sequence. The cycles in the

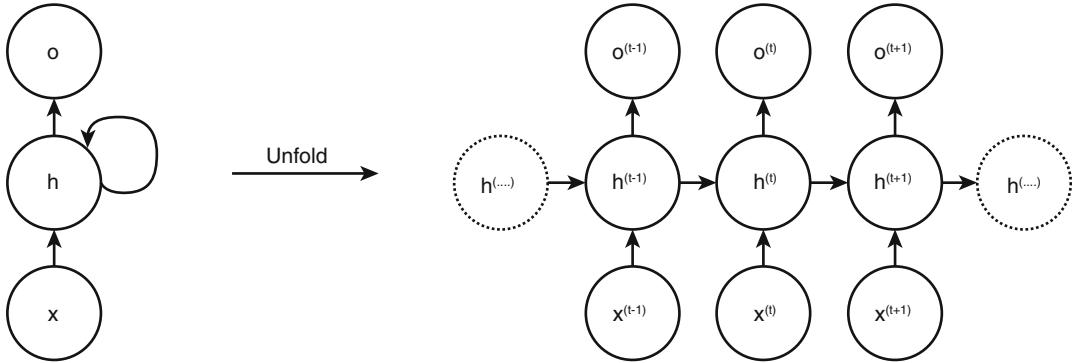


Fig. 1 (Left) Circuit diagram for SimpleRNN with input x being incorporated into hidden state h with a feedback connection and an output o . (Right) The same SimpleRNN network shown as an unfolded computational graph with nodes at every time step

computational graph represent the impact of the past value of a variable on the present time step. The computational graph has a repetitive structure that unfolds the recursive computation of the RNN which corresponds to a chain of events. It shows the flow of the information, forward in the time of computing the outputs and losses and backward when computing the gradients. The unfolded computational graph is shown in Fig. 1. The equation corresponding to the computational graph is $b^{(t)} = f(b^{(t-1)}, x^{(t)}; W)$, where b is the hidden state of the network, x is the input, t is the time step, and W denotes the weights of the network connections comprising of input-to-hidden, hidden-to-hidden, and hidden-to-output connection weights.

2.1.1 Training Fundamentals

Training is performed by gradient computation of the loss function with respect to the parameters involved in forward propagation from left to right of the unrolled graph followed by back-propagation moving from right to left through the graph. Such gradient computation is an expensive operation as the runtime cannot be reduced by parallelism because the forward propagation is sequential in nature. The states computed in the forward pass are stored until they are reused in the back-propagation. The back-propagation algorithm applied to RNN is known as **back-propagation through time** (BPTT) [4].

The following computational operations are performed in RNN during the forward propagation to calculate the output and the loss.

$$\begin{aligned} a^{(t)} &= b + Wh^{(t-1)} + Ux^{(t)} \\ h^{(t)} &= \tanh(a^{(t)}) \\ o^{(t)} &= c + Vh^{(t)} \\ \hat{y}^{(t)} &= \sigma(o^{(t)}) \end{aligned}$$

where b and c are the biases and \mathbf{U} , \mathbf{V} , and \mathbf{W} are the weight matrix for input-to-hidden connections, hidden-to-output connection, and hidden-to-hidden connections respectively, and σ is a sigmoid function. The total loss for a sequence of \mathbf{x} values and its corresponding \mathbf{y} values is obtained by summing up the losses over all time steps.

$$\sum_{t=1}^{\tau} L^{(t)} = L\left((\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\tau)}), (\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(\tau)})\right)$$

To minimize the loss, the gradient of the loss function is calculated with respect to the parameters associated with it. The parameters associated with the nodes of the computational graph are \mathbf{U} , \mathbf{V} , \mathbf{W} , b , c , $\mathbf{x}^{(t)}$, $\mathbf{b}^{(t)}$, $\mathbf{o}^{(t)}$, and $L^{(t)}$. The output $\mathbf{o}^{(t)}$ is the argument to the softmax to obtain the vector $\hat{\mathbf{y}}$ of probabilities over the output. During back-propagation, the gradient for each node is calculated recursively starting with the nodes preceding the final loss. It is then iterated backward in time to back-propagate gradients through time. \tanh is a popular choice for activation function as it tends to avoid vanishing gradient problem by retaining non-zero value longer through the back-propagation process.

2.1.2 SimpleRNN Architecture Variations Based on Parameter Sharing

- *Connection between hidden units.* The RNN produces outputs at every time step, and the parameters are passed between hidden-to-hidden units (Fig. 2a). This corresponds to the standard SimpleRNN presented above and is widely used.
- *Connection between outputs to hidden units.* The RNN produces outputs at every time step, and the parameters are passed from an output at a particular time step to the hidden unit at the next time step (Fig. 2b).
- *Sequential input to single output.* The RNN produces a single output at the end after reading the entire sequence and has connections between the hidden units at every time step (Fig. 2c).

2.1.3 SimpleRNN Architecture Variations Based on Inputs and Outputs

Different variations also exist depending on the number of inputs and outputs:

- *One-to-one:* The traditional RNN has one-to-one input to output mapping at each time step t as shown in Fig. 3a.
- *One-to-many:* One-to-many RNN has one input at a time step for which it generates a sequence of outputs at consecutive time steps as shown in Fig. 3b. This type of RNN architecture is often used for image captioning.

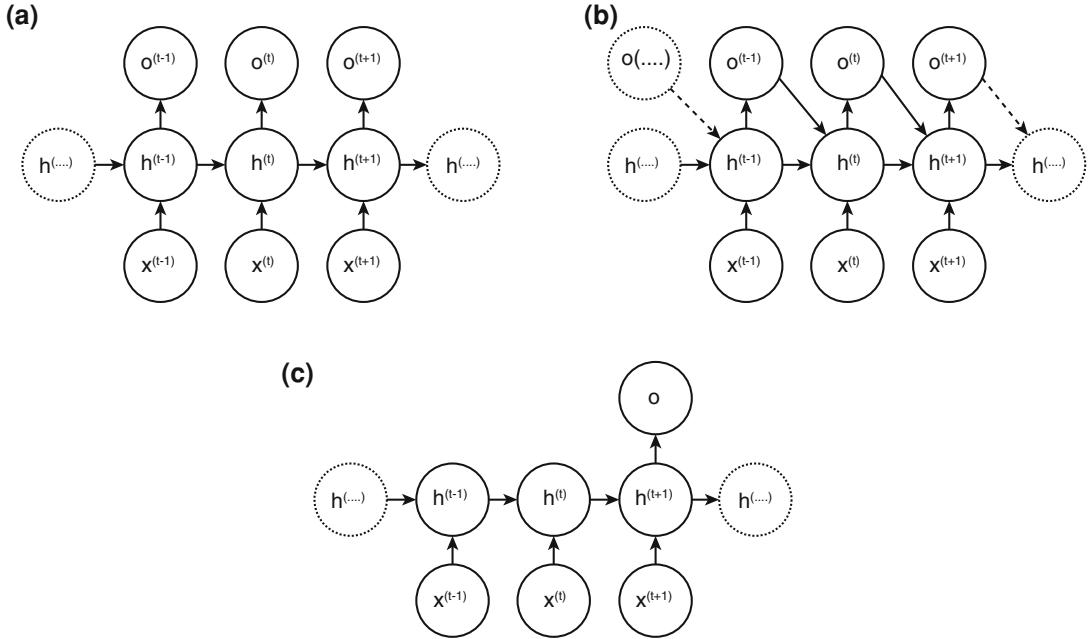


Fig. 2 Types of SimpleRNN architectures based on parameter sharing: (a) SimpleRNN with connections between hidden units, (b) SimpleRNN with connections from output to hidden units, and (c) SimpleRNN with connections between hidden units that read the entire sequence and produce a single output

- *Many-to-one:* Many-to-one RNN has many inputs and one output, at each time step as shown in Fig. 3c. This type of RNN architecture is used for text classification.
- *Many-to-many:* Many-to-many RNN architecture can be designed in two ways. First, the input is taken by the RNN and the corresponding output is given at the same time step as illustrated in Fig. 3d. This type of RNN is used for named entity recognition. Second, the input is taken by the RNN at each time step and the output is given by the RNN at the next time step depending upon all the input sequence as illustrated in Fig. 3e. Popular uses of this type of RNN architecture are in machine translation.

2.1.4 Challenges of Long-Term Dependencies in SimpleRNN

SimpleRNN works well with the short-term dependencies, but when it comes to long-term dependencies, it fails to remember the long-term information. This problem arises due to the vanishing gradient or exploding gradient [6]. When the gradients are propagated over many stages, it tends to vanish most of the times or sometimes explodes. The difficulty arises due to the exponentially smaller weight assigned to the long-term interactions compared to the short-term interactions. It takes very long time to learn the long-term dependencies as the signals from these dependencies tend to be hidden by the small fluctuations arising from the short-term dependencies.

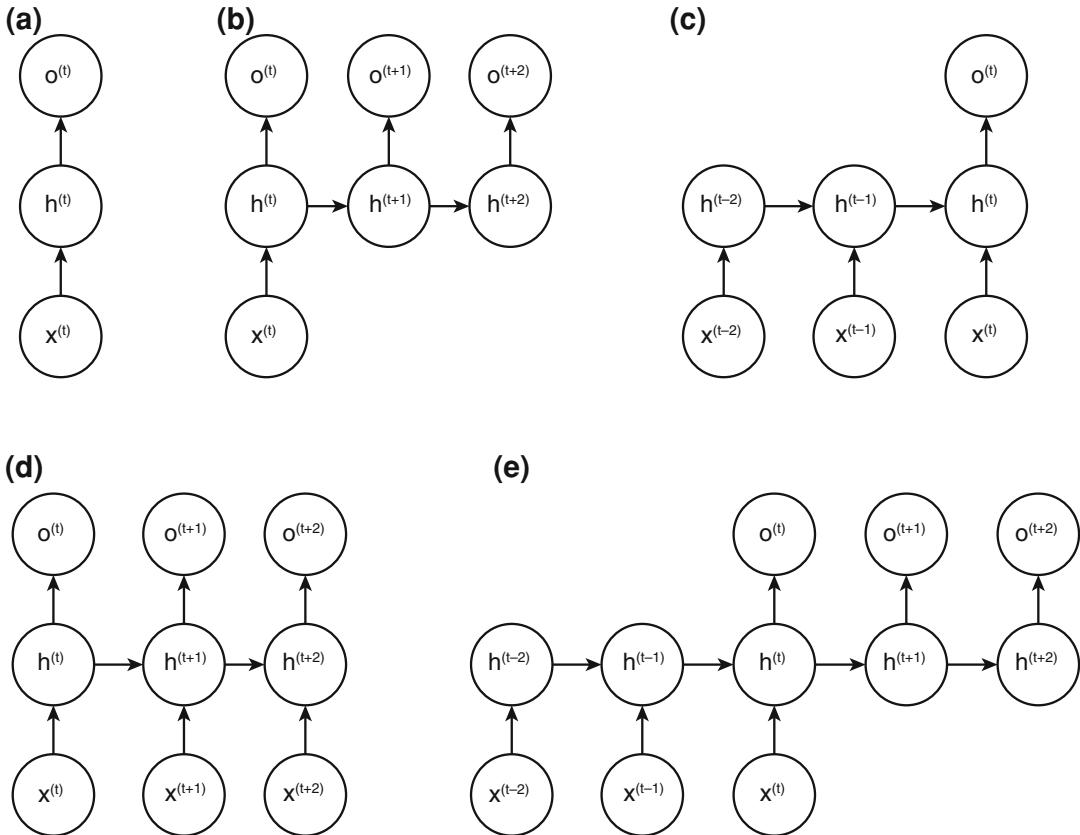


Fig. 3 (a) One-to-one RNN. (b) One-to-many RNN. (c) Many-to-one RNN. (d) Many-to-many RNN. (e) Many-to-many RNN. x represents the input and o represents the output

2.2 Long Short-Term Memory (LSTM)

To address this long-term dependency problem, gated RNNs were proposed. Long short-term memory (LSTM) is a type of gated RNN which was proposed in 1997 [7]. Due to the property of remembering the long-term dependencies, LSTM has been a successful model in many applications like speech recognition, machine translation, image captioning, etc. LSTM has an inner self loop in addition to the outer recurrence of the RNN. The gradients in the inner loop can flow for longer duration and are conditioned on the context rather than being fixed. In each cell, the input and output is the same as that of ordinary RNN but has a system of gating units to control the flow of information. Figure 4 shows the flow of the information in LSTM with its gating units.

There are three gates in the LSTM—the **external input gate**, the **forget gate**, and the **output gate**. The **forget gate** at time t and state s_i ($f_i^{(t)}$) decides which information should be removed from the cell state. The gate controls the self loop by setting the weight between 0 and 1 via a sigmoid function σ . When the value is near to 1, the information of the past is retained, and if the value is near to

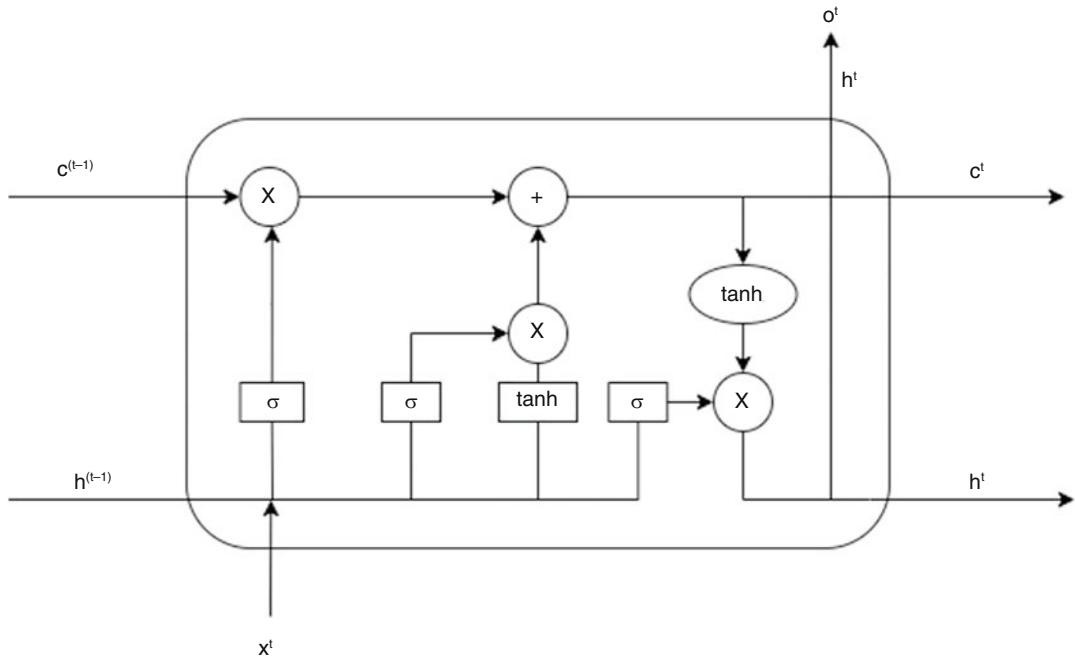


Fig. 4 Long short-term memory with cell state \mathbf{c}^t , hidden state \mathbf{h}^t , input \mathbf{x}^t , and output \mathbf{o}^t

0, the information is discarded. After the **forget gate**, the internal state $s_i^{(t)}$ is updated. Computation for **external input gate** (g_i^t) is similar to that of **forget gate** with a sigmoid function to obtain a value between 0 and 1 but with its own parameters. The **output gate** of the LSTM also has a sigmoid unit which determines whether to output the value or to shut off the value h_i^t via the **output gate** q_i^t .

$$\begin{aligned}
 f_i^{(t)} &= \sigma \left(\sum_j \mathbf{U}_{i,j}^f \mathbf{x}_j^t + \sum_j \mathbf{W}_{i,j}^f \mathbf{h}_j^{(t-1)} + b_i^f \right) \\
 s_i^{(t)} &= f_i^{(t)} s_i^{(t-1)} + g_i^t \sigma \left(b_i + \sum_j \mathbf{U}_{i,j} \mathbf{x}_j^t + \sum_j \mathbf{W}_{i,j} \mathbf{h}_j^{(t-1)} \right) \\
 g_i^t &= \sigma \left(b_i^g + \sum_j \mathbf{U}_{i,j}^g \mathbf{x}_j^t + \sum_j \mathbf{W}_{i,j}^g \mathbf{h}_j^{(t-1)} \right) \\
 h_i^t &= \tanh(s_i^t) q_i^t \\
 q_i^t &= \sigma \left(b_i^o + \sum_j \mathbf{U}_{i,j}^o \mathbf{x}_j^t + \sum_j \mathbf{W}_{i,j}^o \mathbf{h}_j^{(t-1)} \right)
 \end{aligned}$$

\mathbf{x}^t is the input vector at time t , $\mathbf{h}^{(t)}$ is the hidden layer vector, b_i denote the biases, and \mathbf{U}_i and \mathbf{W}_i represent the input weights and the recurrent weights, respectively.

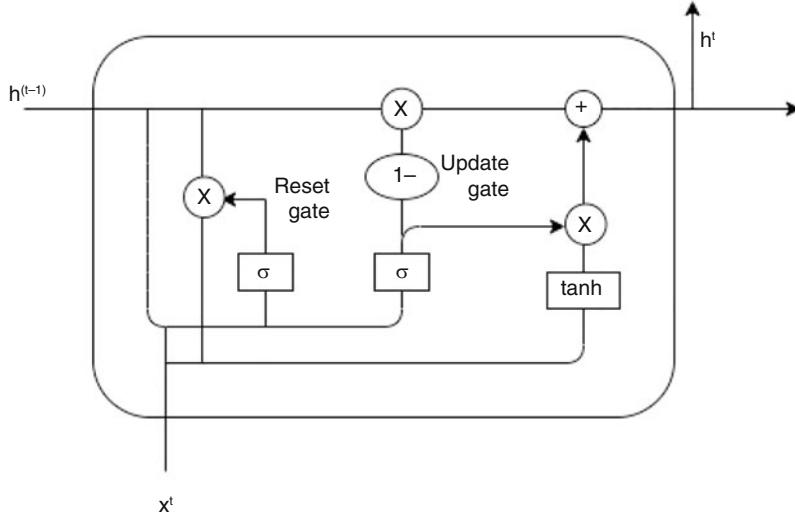


Fig. 5 Gated recurrent neural network (GRU) with input x^t and hidden unit h^t

2.3 Gated Recurrent Unit (GRU)

In LSTM, the computation time is large as there are a lot of parameters involved during back-propagation. To reduce the computation time, gated recurrent unit (GRU) was proposed in the year 2014 by Cho et al. with less gates than in LSTM [8]. The functionality of the GRU is similar to that of LSTM but with a modified architecture. The representation diagram for GRU can be found in Fig. 5. Like LSTM, GRU also solves the vanishing and exploding gradient problem by capturing the long-term dependencies with the help of gating units. There are two gates in GRU, the **reset gate** and the **update gate**. The **reset gate** determines how much of the past information it needs to forget, and the **update gate** determines how much of the past information it needs to carry forward.

The computation at the **reset gate** (r_i^t) and the **update gate** (u_i^t), as well as hidden state (b_i^t) and the at time t , can be represented by the following:

$$\begin{aligned} r_i^{(t)} &= \sigma(b_i^r + \sum_j U_{i,j}^r x_j^{(t)} + \sum_j W_{i,j}^r b_j^{(t)}) \\ u_i^{(t)} &= \sigma(b_i^u + \sum_j U_{i,j}^u x_j^{(t)} + \sum_j W_{i,j}^u b_j^{(t)}) \\ b_i^{(t)} &= u_i^{(t-1)} b_i^{(t-1)} + (1 - u_i) \\ &\quad \times \sigma(b_i + \sum_j U_{i,j} x_j^{(t-1)} + \sum_j W_{i,j} r_j^{(t-1)} b_j^{(t-1)}) \end{aligned}$$

where b_i denotes biases and U_i and W_i denote initial and recurrent weights, respectively.

When the **reset gate** value is close to 0, the previous hidden state value is discarded and reset with the present value. This enables the hidden state to forget the past information that is irrelevant for future. The **update gate** determines how much of the relevant past information to carry forward for future.

The property of the **update gate** to carry forward the past information allows it to remember the long-term dependencies. For short-term dependencies, the **reset gate** will be frequently active to reset with current values and remove the previous ones, while, for long-term dependencies, the **update gate** will be often active for carrying forward the previous information.

2.3.1 Advantage of LSTM and GRU over SimpleRNN

The LSTM and GRU can handle the vanishing gradient issue of SimpleRNN with the help of gating units. The LSTM and GRU have the additive feature that they retain the past information by adding the relevant past information to the present state. This additive property makes it possible to remember a specific feature in the input for longer time. In SimpleRNN, the past information loses its relevance when new input is seen. In LSTM and GRU, any important feature is not overwritten by new information. Instead, it is added along with the new information.

2.3.2 Differences Between LSTM and GRU

There are a few differences between LSTM and GRU in terms of gating mechanism which in turn result in differences observed in the content generated. In LSTM unit, the amount of the memory content to be used by other units of the network is regulated by the **output gate**, whereas in GRU, the full content that is generated is exposed to other units. Another difference is that the LSTM computes the new memory content without controlling the amount of previous state information flowing. Instead, it controls the new memory content that is to be added to the network. On the other hand, the GRU controls the flow of the past information when computing the new candidate without controlling the candidate activation.

2.4 Bidirectional RNN (BRNN)

In SimpleRNN, the output of a state at time t only depends on the information of the past $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t-1)}$ and the present input $\mathbf{x}^{(t)}$. However, for many sequence-to-sequence applications, the present state output depends on the whole sequence information. For example, in language translation, the correct interpretation of the current word depends on the past words as well as the next words. To overcome this limitation of SimpleRNN, bidirectional RNN (BRNN) was proposed by Schuster and Paliwal in the year 1997 [9].

Bidirectional RNNs combine an RNN which moves forward with time, beginning from the start of the sequence, with another RNN that moves backward through time, beginning from the end of the sequence. Figure 6 illustrates a bidirectional RNN with $\mathbf{h}^{(t)}$

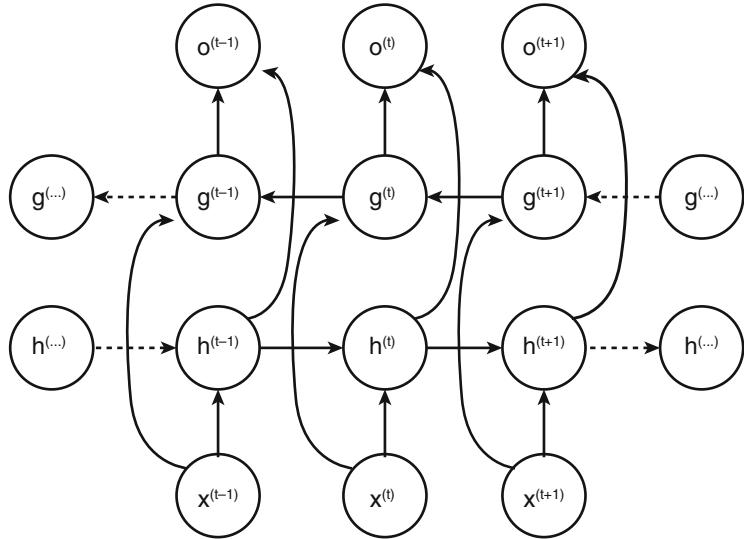


Fig. 6 Bidirectional RNN with forward sub-RNN having h^t hidden state and backward sub-RNN having g^t hidden state

the state of the sub-RNN that moves forward through time and $g^{(t)}$ the state of the sub-RNN that moves backward with time. The output of the sub-RNN that moves forward is not connected to the inputs of sub-RNN that moves backward and vice versa. The output $o^{(t)}$ depends on both past and future sequence data but is sensitive to the input values around t .

2.5 Deep RNN

Deep models are more efficient than their shallow counterparts, and, with the same hypothesis, deep RNN was proposed by Pascanu et al. in 2014 [10]. In “shallow” RNN, there are generally three blocks for computation of parameters: the input state, the hidden state, and the output state. These blocks are associated with a single weight matrix corresponding to a shallow transformation which can be represented by a single-layer multilayer perceptron (MLP). In deep RNN, the state of the RNN can be decomposed into multiple layers. Figure 7 shows in general a deep RNN with multiple deep MLPs. However, different types of depth in an RNN can be considered separately like input-to-hidden, hidden-to-hidden, and hidden-to-output layer. The lower layer in the hierarchy can transform the input into an appropriate representation for higher levels of hidden state. In hidden-to-hidden state, it can be constructed with a previous hidden state and a new input. This introduces additional non-linearity in the architecture which becomes easier to quickly adapt changing modes of the input. By introducing deep MLP in hidden-to-output state makes the layer compact which helps in summarizing the previous inputs and helps in predicting the output easily. Due to the deep MLP in the RNN architecture, the learning becomes slow and optimization is difficult.

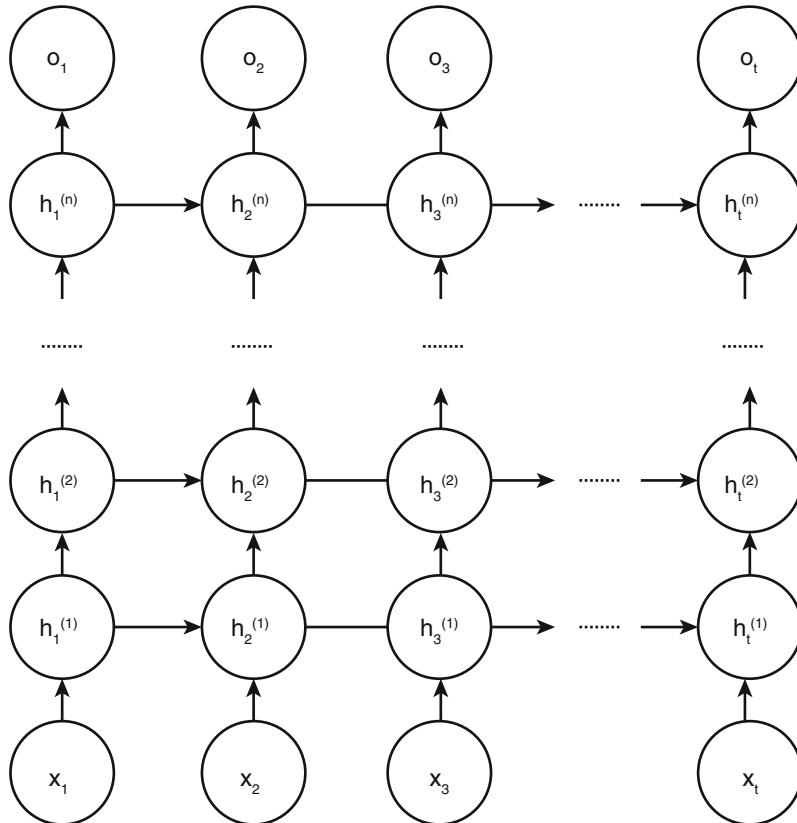


Fig. 7 Deep recurrent neural network

2.6 Encoder–Decoder

Encoder–decoder architecture was proposed by Cho et al. (2014) [8] to map a variable length input sequence to a variable length output sequence. Therefore, it is also known as sequence-to-sequence architecture. Before encoder–decoder was introduced, there were RNN models which were used for sequence-to-sequence applications, but they had limitations as the input and output sequences had to have the same length. Encoder–decoder was used for addressing variable length sequence-to-sequence problems such as machine translation or speech recognition where the input sequence and output sequence lengths may not be the same in most of the cases. Encoder and decoder are both RNNs where the encoder RNN encodes the whole input $\mathbf{X} = \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n_x)}$ into a context vector \mathbf{c} and outputs the context vector \mathbf{c} which is fed as an input to the decoder RNN. The decoder RNN generates an output sequence $\mathbf{Y} = \mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n_y)}$. In the encoder–decoder model, the input length n_x and the output length n_y can be different unlike the previous RNN models. The number of hidden layers in encoder and decoder are not necessarily be the same. The limitation of this architecture is that it fails to properly summarize a

long sequence if the context vector is too small. This problem was solved by Bahdanau et al. (2015) [11] by making the context vector a variable length sequence with added attention mechanism.

2.7 Attention Models (Transformers)

Due to the sequential learning mechanism, the context vector generated by the encoder (*see* Subheading 2.6) is more focused on the later part of the sequence than on the earlier part. An extension to the encoder–decoder model was proposed by Bahdanau et al. [11] for machine translation where the model generates each word based on the most relevant information in the source sentence and previously generated words. Unlike the previous encoder–decoder model where the whole input sequence is encoded into a single context vector, this extended encoder–decoder model learns to give attention to the relevant words present in the source sequence regardless of the position in the sequence by encoding the input sequence into sequences of vectors and chooses selectively while decoding each word. This mechanism of paying attention to the relevant information that are related to each word is known as attention mechanism.

Although this model solves the problem for fixed-length context vectors, the sequential decoding problem still persists. To decode the sequence in less time by introducing parallelism, self-attention was proposed by Google Brain team, Ashish Vaswani et al. [12]. They invented the Transformer model which is based on self-attention mechanism and was designed to reduce the computation time. It computes the representation of a sequence that relates to different positions of the same sequence. The self-attention mechanism was embedded in the Transformer model. The Transformer model has a stack of six identical layers each for encoding the sequence and decoding the sequence as illustrated in Fig. 8. Each layer of the encoder and decoder has sub-layers comprising multi-head self-attention mechanisms and position-wise fully connected layers. There is a residual connection around the two sub-layers followed by normalization. In addition to the two sub-layers, there is a third layer in the decoder that performs multi-head attention over the output of the encoder stack. In the decoder, the multi-head attention is masked to prevent the position from attending the later part of the sequence. This ensures that the prediction for a position p depends only on the positions less than p in the sequence. The attention function can be described as mapping a query and key-value pairs to an output. All the parameters involved in the computation are all vectors. To calculate the output, scalar dot product operation is performed on the query and all keys, and divide each key by $\sqrt{d_k}$ (where d_k is the dimension on the keys). Finally, the softmax is applied to it to obtain the weights on the values. The computation of attention function can be represented by the following equation:

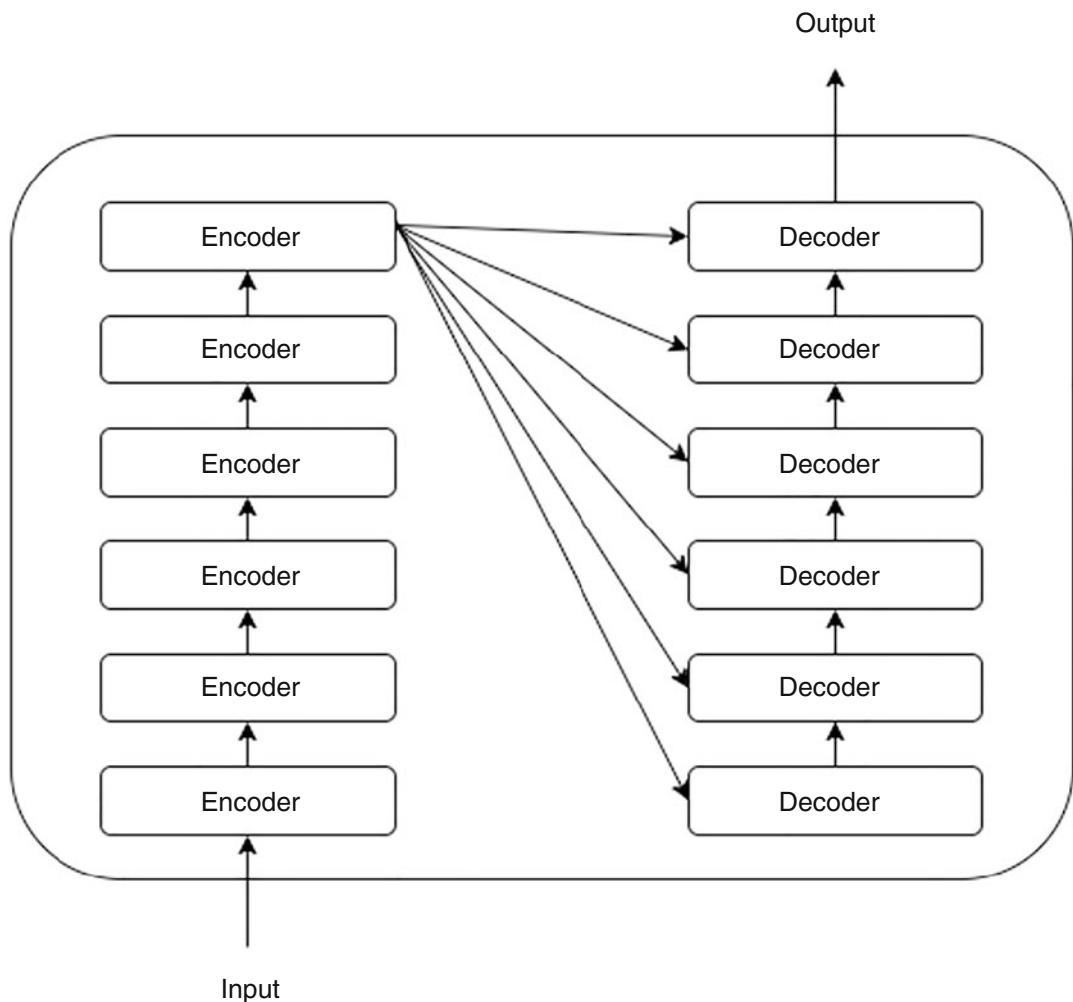


Fig. 8 Transformer with six layers of encoders and six layers of decoders

$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$, where \mathbf{Q} , \mathbf{K} , and \mathbf{V} are all matrices corresponding to query, keys, and values, respectively. A more in-depth coverage of Transformers is provided in Chap. 6.

3 Tips and Tricks for RNN Training

As previously stated, the vanishing gradient and exploding gradient problems are well-known concerns when it comes to properly training RNN models [13, 14]. The fundamental challenge arises from the fact that RNNs can be naturally unfolded, allowing their recurrent connections to perform feedforward calculations, which result in an RNN with the same number of layers as the number of elements in the sequence. Two major issues arise as a result:

- *Gradient vanishing problem.* It becomes difficult to effectively learn long-term dependencies in sequences due to the gradient vanishing problem [6]. As a result, a prospective model prediction will be essentially unaffected by earlier layers.
- *Exploding gradient problem.* Adding more layers to the network amplifies the effect of large gradients, increasing the risk of a learning derailment since significant changes to the network weights can be performed at each step, potentially causing the gradients to blow out exponentially. In fact, weights that are closer to the input layer will obtain larger updates than weights that are closer to the output layer, and the network may become unable to learn correlations between temporally distant events.

To overcome these limitations, we need to create solutions so that the RNN model can work on various time scales, with some sections operating on fine-grained time scales and handling small details and others operating on coarse time scales and efficiently transferring information from the distant past to the present. In this section, we discuss several popular strategies to tackle these issues.

3.1 Skip Connection

The practice of skipping layers effectively simplifies the network by using fewer direct connected layers in the initial training stages. This speeds learning by reducing the impact of vanishing gradients, as there are fewer layers to propagate through. As the network learns the feature space during the training phase, it gradually restores the skipped layers. Lin et al. [15] proposed the use of such skip connections, which follows from the idea of incorporating delays in feedforward neural networks from Lang et al. [16]. Conceptually, skip connections are a standard module in deep architectures and are commonly referred to as residual networks, as described by He et al. [17]. They are responsible to skip layers in the neural network and feeding the output of one layer as the input to the next layers. This technique is used to allow gradients to flow through a network directly, without passing through non-linear activation functions, and it has been empirically proven that these additional steps are often beneficial for the model convergence [17]. Skip connections can be used through the non-sequential layer in two fundamental ways in neural networks:

- **Additive Skip Connections.** In this type of design, the data from early layers is transported to deeper layers via matrix addition, causing back-propagation to be done via addition (Fig. 9b). This procedure does not require any additional parameters because the output from the previous layer is added to the layer ahead. One of the most common techniques used in this type of architecture is to stack the skip residual blocks together and use an identity function to preserve the gradient [18]. The core concept is to use a vector addition to back-

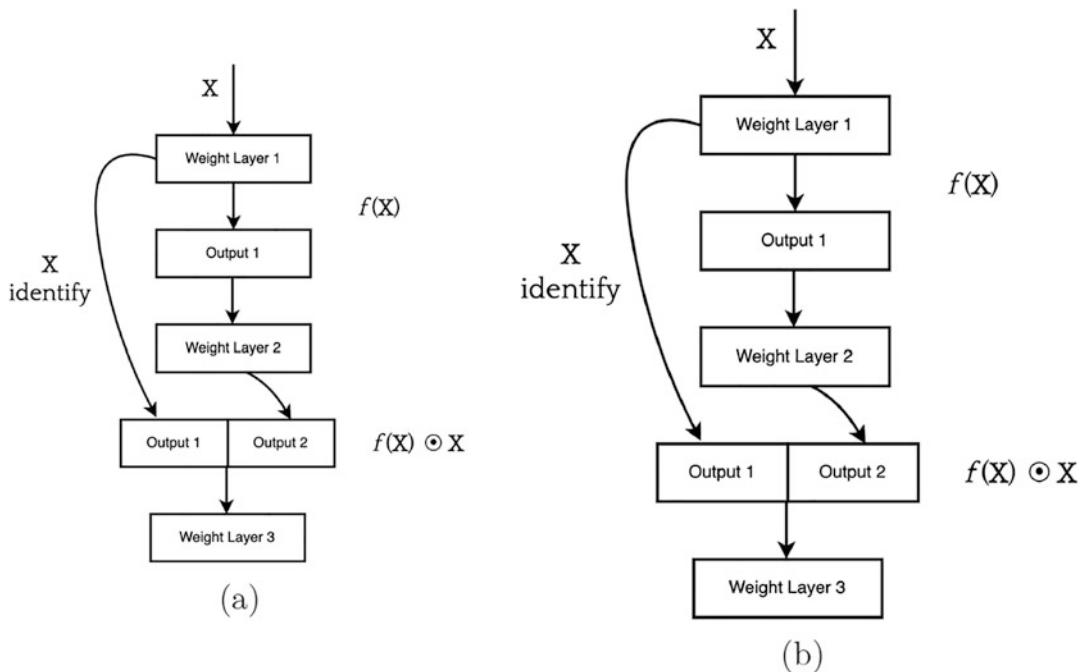


Fig. 9 Skip connection residual architectures: **(a)** concatenate output of previous layer and skip connection; **(b)** sum of the output of previous layer and skip connection

propagate through the identity function. The gradient is then simply multiplied by one, and its value is preserved in the earlier layers.

- **Concatenative Skip Connections.** Another way for establishing skip connections is to concatenate previous feature maps. The aim of concatenation is to leverage characteristics acquired in prior layers to deeper layers. In addition, concatenating skip connections provides an alternate strategy for assuring feature reusability of the same dimensionality from prior layers without the need to learn duplicate maps. Figure 9(a) illustrates a diagram example of how the architecture looks like. The primary concept of the architecture is to allow subsequent layers to reuse intermediary representations, allowing them to maintain more information and enhance long-term dependency performance.

3.2 Leaky Units

One of the major challenges when training RNNs is capturing long-term dependencies and efficiently transferring information from distant past to present. An effective method to obtain coarse time scales is to employ leaky units [19], which are hidden units with linear self-connections and a weight on the connections that is close to one. In a leaky RNN, hidden units are able to access values from prior states and can be utilized to obtain temporal representations. Formula $b_t = \alpha * b_{t-1} + (1 - \alpha) * h_t$ expresses the state update

rule of a leaky unit, where $\alpha \in (0, 1)$ is an example of a linear self-connection from \mathbf{h}_{t-1} to \mathbf{h}_t , and it is a parameter to be learned during the training stage. Essentially, α controls the information flow in the state. When α is near one, the state is almost unchanged, and information about the past is retained for a long time, and when α is close to zero, the information about the past is rapidly discarded, and the state is largely replaced by a new state \mathbf{h}_t .

3.3 Clipping Gradients

Gradient clipping is a technique that tries to overcome the exploding gradient problem in RNN training, by constraining gradient norms (element-wise) to a predetermined minimum or maximum threshold value since the exploding gradients are clipped and the optimization begins to converge to the minimum point. Gradient clipping can be used in two fundamental ways:

- **Clipping-by-value.** Using this technique, we define a minimum clip value and a maximum clip value. If a gradient exceeds the threshold value, we clip the gradient to the maximum threshold. If the gradient is less than the lower limit of the threshold, we clip the gradient to the minimum threshold.
- **Clipping-by-norm.** The idea behind this technique is very similar to clipping-by-value. The key difference is that we clip the gradients by multiplying the unit vector of the gradients with the threshold. Gradient descent will be able to behave properly even if the loss landscape of the model is irregular since the weight updates will also be rescaled. This significantly reduces the likelihood of an overflow or underflow of the model.

4 RNN Applications in Language Modeling

Language modeling is the process of learning meaningful vector representations for language or text using sequence information and is generally trained to predict the next token or word given the input sequence of tokens or words. Bengio et al. [20] proposed a framework for neural network-based language modeling. RNN architecture is particularly suited to processing free-flowing natural language due to its sequential nature. As described by Mikolov et al. [21], RNNs can learn to compress a whole sequence as opposed to feedforward neural networks that compress only a single input item. Language modeling can be an independent task or be part of a language processing pipeline with downstream prediction or classification task. In this section, we will discuss applications of RNN for various language processing tasks.

4.1 Text Classification

Many interesting real-world applications concerning language data can be modeled as text classification. Examples include sentiment classification, topic or author identification, and spam detection with applications ranging from marketing to query-answering [22, 23]. In general, models for text classification include some RNN layers to process sequential input text [22, 23]. The embedding of the input learnt by these layers is later processed through varying classification layers to predict the final class label. Many-to-one RNN architectures are often employed for text classification.

As a recent technical innovation, RNNs have been combined with convolutional neural networks (CNNs), thus combining the strengths of two architectures, to process textual data for classification tasks. LSTMs are popular RNN architecture for processing textual data because of their ability to track patterns over long sequences, while CNNs have the ability to learn spatial patterns from data with two or more dimensions. Convolutional LSTM (C-LSTM) combines these two architectures to form a powerful architecture that can learn local phrase-level patterns as well as global sentence-level patterns [24]. While CNN can learn local and position-invariant features and RNN is good at learning global patterns, another variation of RNN has been proposed to introduce position-invariant local feature learning into RNN. This variation is called disconnected RNN (DRNN) [25]. Information flow between tokens/words at the hidden layer is limited by a hyper-parameter called *window size*, allowing the developer to choose the *width* of the *context* to be considered while processing text. This architecture has shown better performance than both RNN and CNN on several text classification tasks [25].

4.2 Text Summarization

Text summarization approaches can be broadly categorized into (1) extractive and (2) abstractive summarization. The first approach relies on selection or extraction of sentences that will be part of the summary, while the latter generates new text to build a summary. RNN architectures have been used for both types of summarization techniques.

4.2.1 Extractive Text Summarization

Extractive summarization frameworks use many-to-one RNN as a classifier to distinguish sentences that should be part of the summary. For example, a two-layer RNN architecture is presented in [26] where one layer processes words in one sentence and the other layer processes many sentences as a sequence. The model generates sentence-level labels indicating whether the sentence should be part of the summary or not, thus producing an extractive summary of the input document. Xu et al. have presented a more sophisticated extractive summarization model that not only extracts sentences to be part of the summary but also proposes possible syntactic compressions for those sentences [27]. Their proposed architecture is a

combination of CNN and bidirectional LSTM, while a neural classifier evaluates possible syntactic compressions in the context of the sentence as well as the broader context of the document.

4.2.2 Abstractive Text Summarization

Abstractive summarization frameworks expect the RNN to process input text and generate a new sequence of text that is the summary of input text, effectively using many-to-many RNN as a text generation model. While it is relatively straightforward for extractive summarizers to achieve basic grammatical correctness as correct sentences are picked from the document to generate a summary, it has been a major challenge for abstractive summarizers. Grammatical correctness depends on the quality of the text generation module. Grammatical correctness of abstractive text summarizers has improved recently due to developments in contextual text processing, language modeling, as well as availability of computational power to process large amounts of text.

Handling of rare tokens/words is a major concern for modern abstractive summarizers. For example, proper nouns such as specific names of people and places occur less frequently in the text; however, generated summaries are incomplete and incomprehensible if such tokens are ignored. Nallapati et al. proposed a novel solution composed of GRU-RNN layers with attention mechanism by including switching decoder in their abstractive summarizer architecture [28] where the text generator module has a switch which can enable the module to choose between two options: (1) generate a word from the vocabulary and (2) point to one of the words in the input text. Their model is capable of handling rare tokens by pointing to their position in the original text. They also employed *large vocabulary trick* which limits the vocabulary of the generator module to tokens of the source text only and then adds frequent tokens to the vocabulary set until its size reaches a certain threshold. This trick is useful in limiting the size of the network.

Summaries have latent structural information, i.e., they convey information following certain linguistic structures such as “What-Happened” or “Who-Action-What.” Li et al. presented a recurrent generative decoder based on variational auto-encoder (VAE) [29]. VAE is a generative model that takes into account latent variables, but is not inherently sequential in nature. With the historical dependencies in latent space, it can be transformed into a sequential model where generative output is taking into account history of latent variables, hence producing a summary following latent structures.

4.3 Machine Translation

Neural machine translation (NMT) models are trained to process input sequence of text and generate an output sequence which is the translation of the input sequence in another language. As mentioned in Subheading 2.6, machine translation is a classic example of conversion of one sequence to another using encoder–

decoder architecture where lengths of both sequences may be different. In 2014, many-to-many RNN-based encoder–decoder architecture was proposed where one RNN encodes the input sequence of text to a fixed-length vector representation, while another RNN decodes the fixed-length vector to the target translated sequence [30]. Both RNNs are jointly trained to maximize the conditional probability of the target sequence given the input sequence. Later, attention-based modeling was added to vanilla encoder–decoder architecture for machine translation. Luong et al. discussed two types of attention mechanism in their work on NMT: (i) global and (ii) local attention [31]. In global attention, a global context vector is estimated by learning variable length alignment and attention scores for all source words. In local attention, the model predicts a single aligned position for the current target word and then computes a local context vector from attention predicted for source words within a small window of the aligned position. Their experiments show significant improvement in translation performance over models without attention. Local attention mechanism has the advantage of being computationally less expensive than global attention mechanism.

4.4 Image-to-Text Translation

Image-to-text translation models are expected to convert visual data (i.e., images) into textual data (i.e., words). In general, the image input is passed through some convolutional layers to generate a dense representation of the visual data. Then, the embedded representation of the visual data is fed to an RNN to generate a sequence of text. Many-to-one RNN architectures are popular for this task.

In 2015, Karpathy et al. [32] presented their influential work on training region convolutional neural network (RCNN) to generate representation vectors for image regions and bidirectional RNN to generate representation vectors for corresponding caption in semantic alignment with each other. They also proposed novel multi-modal RNN to generate a caption that is semantically aligned with the input image. Image regions were selected based on the ranked output of an object detection CNN.

Xu et al. proposed an attention-based framework to generate image caption that was inspired by machine translation models [33]. They used image representations generated by lower convolutional layers from a CNN model rather than the last fully connected layer and used an LSTM to generate words based on hidden state, last generated word, and context vector. They defined the context vector as a dynamic representation of the image generated by applying an attention mechanism on image representation vectors from lower convolutional layers of CNN. Attention mechanism allowed the model to dynamically select the region to focus on while generating a word for image caption. An additional advantage of their approach was intuitive visualization of the

model's focus for generation of each word. Their visualization experiments showed that their model was focused on the right part of the image while generating each important word.

Such influential works in the field of automatic image captioning were based on image representations generated by CNNs designed for object detection. Some recently proposed captioning models have sought to change this trend. Biten et al. proposed a captioning model for images used to illustrate new articles [34]. Their caption generation LSTM takes into account both CNN-generated image features and semantic embeddings to the text of corresponding new articles to generate a template of a caption. This template contains spaces for the names of entities like organizations and places. These places are filled in using attention mechanism on the text of the corresponding article.

4.5 ChatBot for Mental Health and Autism Spectrum Disorder

ChatBots are automatic conversation tools that have gained vast popularity in e-commerce and as digital personal assistants like Apple's Siri and Amazon's Alexa. ChatBots represent an ideal application for RNN models as conversations with ChatBots represent sequential data. Questions and answers in a conversation should be based on past iterations of questions and answers in that conversation as well as patterns of sequences learned from other conversations in the dataset.

Recently, ChatBots have found application in screening and intervention for mental health disorders such as autism spectrum disorder (ASD). Zhong et al. designed a Chinese-language ChatBot using bidirectional LSTM in sequence-to-sequence framework which showed great potential for conversation-mediated intervention for children with ASD [35]. They used 400,000 selected sentences from chatting histories involving children in many cases. Rakib et al. developed similar sequence-to-sequence model based on Bi-LSTM to design a ChatBot to respond empathetically to mentally ill patients [36]. A detailed survey of medical ChatBots is presented in [37]. This survey includes references to ChatBots built using NLP techniques, knowledge graphs, as well as modern RNN for a variety of applications including diagnosis, searching through medical databases, dialog with patients, etc.

5 Conclusion

Due to the sequential nature of their architecture, RNNs are applied for ordinal or temporal problems, such as language translation, text summarization, and image captioning, and are incorporated into popular applications such as Siri, voice search, and Google Translate. In addition, they are also often used to analyze longitudinal data in medical applications (i.e., cases where repeated observations are available at different time points for each

patient of a dataset). While research in RNN is still an evolving area and new architectures are being proposed, this chapter summarizes fundamentals of RNN including different traditional architectures, training strategies, and influential work. It may serve as a stepping stone for exploring sequential models using RNN and provides reference pointers.

References

1. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by backpropagating errors. *Nature* 323(6088): 533–536
2. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci* 79(8): 2554–2558
3. Schmidhuber J (1993) Netzwerkarchitekturen, Zielfunktionen und Kettenregel (Network architectures, objective functions, and chain rule), Habilitation thesis, Institut für Informatik, Technische Universität München
4. Mozer MC (1995) A focused backpropagation algorithm for temporal. *Backpropag Theory Architect Appl* 137
5. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
6. Hochreiter S (1998) The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int J Uncertainty Fuzziness Knowledge Based Syst* 6(02):107–116
7. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8): 1735–1780
8. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. Preprint. arXiv:14061078
9. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. *IEEE Trans Signal Process* 45(11):2673–2681
10. Pascanu R, Gulcehre C, Cho K, Bengio Y (2013) How to construct deep recurrent neural networks. Preprint. arXiv:13126026
11. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. Preprint. arXiv:14090473
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
13. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw* 5(2):157–166. <https://doi.org/10.1109/72.279181>
14. Pascanu R, Mikolov T, Bengio Y (2013) On the difficulty of training recurrent neural networks. In: Dasgupta S, McAllester D (eds) Proceedings of the 30th international conference on machine learning, PMLR, Atlanta, vol 28, pp 1310–1318
15. Berger AL, Pietra VJD, Pietra SAD (1996) A maximum entropy approach to natural language processing. *Comput Linguist* 22(1): 39–71
16. Becker S, Hinton G (1992) Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature* 355:161–163. <https://doi.org/10.1038/355161a0>
17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778. <https://doi.org/10.1109/CVPR.2016.90>
18. Wu H, Zhang J, Zong C (2016) An empirical exploration of skip connections for sequential tagging. Preprint. arXiv:161003167
19. Jaeger H (2002) Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach. GMD-Forschungszentrum Informationstechnik 5
20. Bengio Y, Ducharme R, Vincent P (2001) A neural probabilistic language model. In: Advances in neural information processing systems, pp 932–938
21. Mikolov T, Karafiat M, Burget L et al (2010) Recurrent neural network based language model. In: INTERSPEECH 2010. Citeseer
22. Jain G, Sharma M, Agarwal B (2019) Optimizing semantic lstm for spam detection. *Int J Inform Technol* 11(2):239–250

23. Bagnall D (2015) Author identification using multi-headed recurrent neural networks. Preprint. arXiv:150604891
24. Zhou C, Sun C, Liu Z, Lau F (2015) A C-LSTM neural network for text classification. Preprint. arXiv:151108630
25. Wang B (2018) Disconnected recurrent neural networks for text categorization. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), pp 2311–2320
26. Nallapati R, Zhai F, Zhou B (2017) Summarunner: a recurrent neural network based sequence model for extractive summarization of documents. In: Thirty-first AAAI conference on artificial intelligence
27. Xu J, Durrett G (2019) Neural extractive text summarization with syntactic compression. Preprint. arXiv:190200863
28. Nallapati R, Zhou B, dos Santos C, Gulcehre Ç, Xiang B (2016) Abstractive text summarization using sequence-to-sequence rnns and beyond. In: Proceedings of the 20th SIGNLL conference on computational natural language learning, pp 280–290
29. Li P, Lam W, Bing L, Wang Z (2017) Deep recurrent generative decoder for abstractive text summarization. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 2091–2100
30. Cho K, van Merriënboer B, Gülcəhre Ç, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: The 2014 conference on empirical methods in natural language processing (EMNLP)
31. Luong MT, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. Preprint. arXiv:150804025
32. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3128–3137
33. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. In: International conference on machine learning, PMLR, pp 2048–2057
34. Biten AF, Gomez L, Rusinol M, Karatzas D (2019) Good news, everyone! context driven entity-aware captioning for news images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12466–12475
35. Zhong H, Li X, Zhang B, Zhang J (2020) A general chinese chatbot based on deep learning and its' application for children with ASD. Int J Mach Learn Comput 10:519–526. <https://doi.org/10.18178/ijmlc.2020.10.4.967>
36. Rakib AB, Rumky EA, Ashraf AJ, Hillas MM, Rahman MA (2021) Mental healthcare chatbot using sequence-to-sequence learning and bilstm. In: Brain informatics, Springer international publishing, pp 378–387
37. Tjiptomongsoguno ARW, Chen A, Sanyoto HM, Irwansyah E, Kanigoro B (2020) Medical chatbot techniques: a review. In: Silhavy R, Silhavy P, Prokopova Z (eds) Software engineering perspectives in intelligent systems. Springer International Publishing, Cham, pp 346–356

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 5

Generative Adversarial Networks and Other Generative Models

Markus Wenzel

Abstract

Generative networks are fundamentally different in their aim and methods compared to CNNs for classification, segmentation, or object detection. They have initially been meant not to be an image analysis tool but to produce naturally looking images. The adversarial training paradigm has been proposed to stabilize generative methods and has proven to be highly successful—though by no means from the first attempt.

This chapter gives a basic introduction into the motivation for generative adversarial networks (GANs) and traces the path of their success by abstracting the basic task and working mechanism and deriving the difficulty of early practical approaches. Methods for a more stable training will be shown, as well as typical signs for poor convergence and their reasons.

Though this chapter focuses on GANs that are meant for image generation and image analysis, the adversarial training paradigm itself is not specific to images and also generalizes to tasks in image analysis. Examples of architectures for image semantic segmentation and abnormality detection will be acclaimed, before contrasting GANs with further generative modeling approaches lately entering the scene. This will allow a contextualized view on the limits but also benefits of GANs.

Key words Generative models, Generative adversarial networks, GAN, CycleGAN, StyleGAN, VQGAN, Diffusion models, Deep learning

1 Introduction

Generative adversarial networks are a type of neural network architecture, in which one network part generates solutions to a task and another part compares and rates the generated solutions against a priori known solutions. While at first glimpse this does not sound much different from any loss function, which essentially also compares a generated solution with the gold standard, there is one fundamental difference. A loss function is static, but the “judge” or “discriminator” network part is trainable (Fig. 1). This means that it can be trained to distinguish the generated from the true solutions and, as long as it succeeds in its task, a training signal for the generative part can be derived. This is how the notion of adversaries came into the name GAN. The discriminator part is

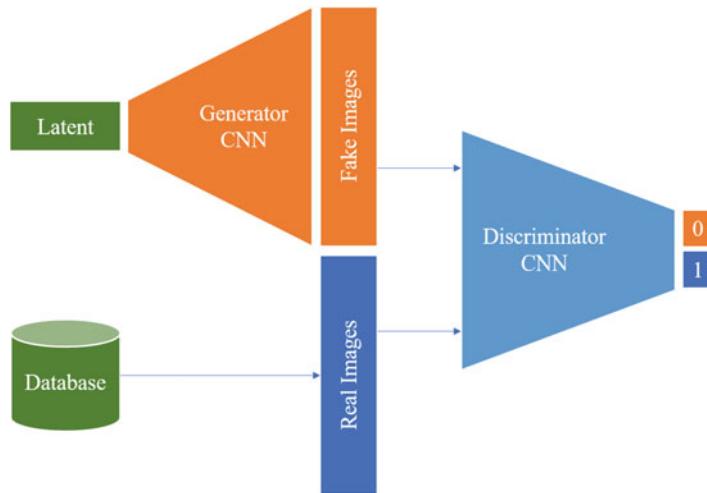


Fig. 1 The fundamental GAN setup for image generation consisting of a generator and a discriminator network; here, CNNs

trained to distinguish true from generated solutions, while the generative part is trained to arrive at the most realistic-looking solutions, making them adversaries with regard to their aims.

Generative adversarial networks are now among the most powerful tools to create naturally looking images from many domains. While they have been created in the context of image generation, the original publication describes the general idea of how to make two networks learn by competing, regardless of the application domain. This key idea can be applied to generative tasks beyond image creation, including text generation, music generation, and many more.

The research interest skyrocketed in the years after the first publication proposing an adversarial training paradigm [1]. Looking at the number of web searches for the topic “generative adversarial networks” shows how the interest in the topic has rapidly grown but also the starting decline of the last years. Authors since 2014 have cast all kinds of problems into the GAN framework, to enable this powerful training mechanism for a variety of tasks, including image analysis tasks as well. This is surprising at first, since there is no immediate similarity between a generative task and, for example, a segmentation or detection task. Still, as evidenced by the success in these application areas, the adversarial training approach can be applied with benefits. Clearly, the decline in interest can to some degree be attributed to the emergence of best practices and proven implementations, while simultaneously the scientific interest has recently shifted to successor approaches. However, similar to the persistent relevance of CNN architectures like ResNets for classification, Mask R-CNNs for detection, or basic transformer architectures for sequence processing, GANs will

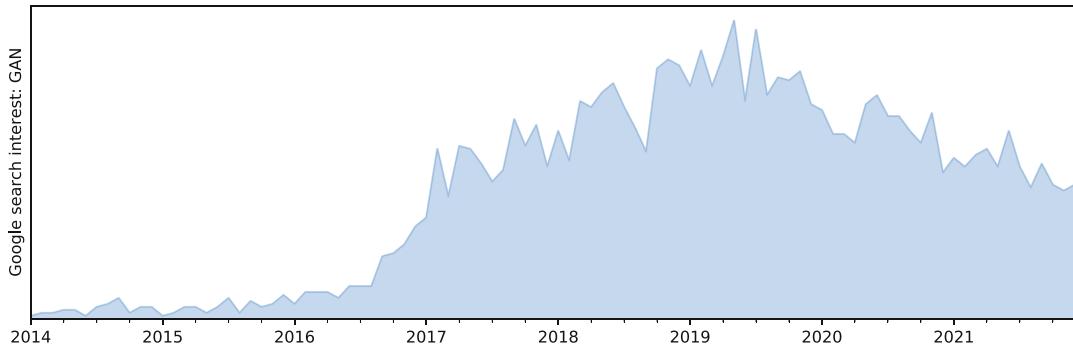


Fig. 2 Google web search-based interest estimate for “generative adversarial networks” since 2014. Relative scale

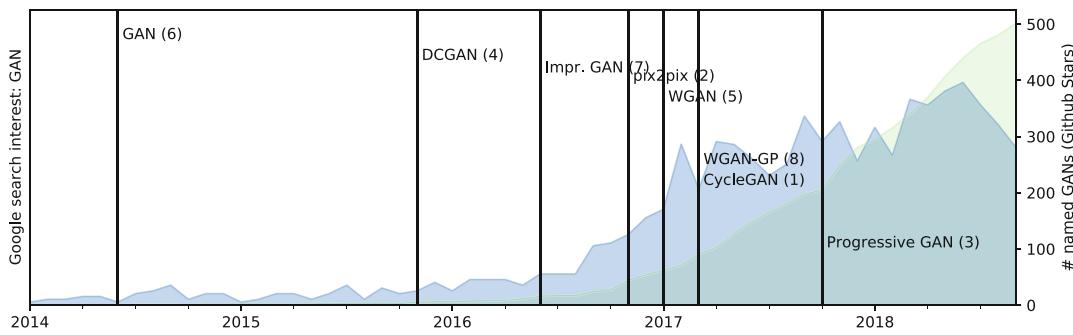


Fig. 3 Some of the most-starred shared GAN code repositories on Github, until 2018. Ranking within this selection in brackets

remain an important tool for image creation and image analysis. The adversarial training paradigm has become an ingredient to models apart from generative aims, providing flexible ways to custom-tailor loss components for given tasks (compare Figs. 2 and 3).

2 Generative Models

Generative processes are fundamentally hard to grasp computationally. Their nature and purpose is to create something “meaningful” out of something less meaningful (even random). The first question to ask therefore is how this can even be possible for a computer program since, intuitively, creation requires an inventive spirit—call it creativity, to use the term humans tend to associate with this. To introduce some of the terminology and basic concepts that we will use in the remainder of this section, some remarks on human creativity will set the scene.

In fact, creative human acts are inherently limited by our concepts of the world, acquired by learning and experience through the

sensory means we have available, and by the available expressive means (tools, instruments, ...) with which we can even conceive of creating something. This is true for any kind of creative act, including writing, painting, wood carving, or any other art, and similarly also for computer programming, algorithm development, or science in general. Our limited internal representation of the world around us frames our creative scope.

This is very comparable to the way computerized, programmed, or learned generative processes create output. They have either an in-build mechanism, or a way to acquire such a mechanism, that represents the tools by which creation is possible, as well as a model of the world that defines the scope of outputs. Practically, a CNN-based generative process uses convolutions as the in-built tool and is by this tool geared to produce image-like outputs. The convolutional layers, if not a priori defined, will represent a set of operations defined by a training process and limited in their expressiveness by the training material—by the fraction of the world that was presented. This will lead us to the fundamental notion of how to capture the variability of the “fraction of the world” that is interesting and how to make a neural network represent this partial world knowledge. It is interesting to note at this point that neither for human creative artists nor for neural networks the ability to (re)create convincing results implies an understanding of the way the templates (in the real world) have come into existence. Generating convincing artifacts does not imply understanding nature. Therefore, GANs cannot explain the parts of nature they are able to generate.

2.1 The Language of Generative Models: Distributions, Density Estimation, and Estimators

Understanding the principles of generative models requires a basic knowledge of distributions. The reason is that—as already hinted at in the previous section—the “fraction of the world” is in fact something that can be thought of as a distribution in a parameter space. If you were to describe a part of the world in a computer-interpretable way, you would define descriptive parameters. To describe persons, you could characterize them by simple measures like age, height, weight, hair and eye color, and many more. You could add blood pressure, heart rate, muscle mass, maximum strength, and more, and even a whole-genome sequencing result might be a parameter. Each of the parameters individually can be collected for the world population, and you will obtain a picture of how this parameter is “distributed” worldwide. In addition, parameters will be in relation with each other, for example, age and maximum strength. Countless such relationships exist, of which the majority are and probably will remain unknown. Those interrelationships are called a joint distribution. Would you know the joint distribution, you could “create” a plausible parameter combination of a nonexisting human. Let us formalize these thoughts now.

2.1.1 Distributions

A distribution describes the frequency of particular observations when watching a random process. Plotting the number of occurrences over an axis of all possible observations creates a histogram. If the possible observations can be arranged on a continuous scale, one can see that observations cluster in certain areas, and we say that they create a “density” or are “dense” there. Hence, when trying to describe where densities are in parameter space, this is associated with the desire to reproduce or sample from distributions, like we want to do it to generate instances from a domain. Before being able to reproduce the function that generates observations, estimating where the dense areas are is required. This will in the most general sense be called density estimation.

Sometimes, the shape of the distribution follows an analytical formula, for example, the normal distribution. If such a closed-form description of the distribution can be given, for instance, the normal distribution, this distribution generalizes the shape of the histogram of observations and makes it possible to produce new observations very easily, by simply sampling from the distribution. When our observations follow a normal distribution, we mean that we expect to observe instances more frequently around the mean of the normal distribution than toward the tails. In addition, the standard deviation quantifies how much more likely observations close to the mean are compared to observations in the tails. We describe our observations with a parametric description of the observed density.

In the remainder of this section, rather than providing a rigorous mathematical definition and description of the mathematics of distributions and (probability) density estimation, we will introduce the basic concepts and terminology in an intuitive way (also compare [Box 1](#)). Readers with the wish for a more in-depth treatment can find tutoring material in the references [\[2–6\]](#).

Box 1: Probability Distributions: Terminology

Several common terms regarding distributions have intuitive interpretations which are given in the following. Let a be an event from the probability distribution A , written as $a \sim A$, and $b \sim B$ an event from another probability distribution.

In a medical example, A might be the distribution of possible neurological diseases and B the distribution of all possible variations of smoking behavior.

Conditional Probability $P(A|B)$ The conditional probability of a certain $a \sim A$, for example, a stroke, might depend on the concrete smoking history of a person,

(continued)

Box 1 (continued)**Joint Probability $P(A, B)$**

described by $b \sim B$. The conditional probability is written as $p(a|b)$ for the concrete instances or $P(A|B)$ if talking about the entire probability distributions A and B .

The probability of seeing instantiations of A and B together is termed the joint probability. Notably, if expanded, this will lead to a large table of probabilities, joining each possible $a \sim A$ (e.g., stroke, dementia, Parkinson's disease, etc.) with each possible $b \sim B$ (casual smoker, frequent smoker, nonsmoker, etc.).

Marginal Probability

The marginal probabilities of A and B (denoted, respectively, $P(A)$ and $P(B)$) are the probabilities of each possible outcome across (and independent of) all of the possible outcomes of the other distribution. For example, it is the probability of seeing nonsmokers across all neurological diseases or seeing a specific disease regardless of smoking status. It is said to be the probability of one distribution marginalized over the other probability distributions.

2.1.2 Density Estimation

We assume in the following that our observations have been produced by a function or process that is not known to us and that cannot be guessed from an arrangement of the observations. In a practical example, the images from a CT or MRI scanner are produced by such a function. Notably, the concern is less about the intractability of the imaging physics but about the appearance of the human body. The imaging physics might be modeled analytically up to a certain error. But the outer shape and inner structure of the

human body and its organs depend on a large amount of mutually influencing factors. Some of these factors are known and can even be modeled, but many are not. In particular, the interdependence of factors must be assumed to be intractable. What we can accumulate is measured data providing information about the body, its shape, and its function. While many measurement instruments exist in medicine, for this chapter, we will be concerned with images as our observations. In the following thought experiment, we will explore a naïve way to model the distribution and try to generate images.

The first step is to examine the gray value distribution or, in other words, estimate the density of values. The most basic way for estimating a density is plotting a histogram. Let the value on the x axis be the image gray value of the medical image in question (in CT expressed in Hounsfield units (HU) and in arbitrary units for MRI). Two plots show histograms of a head MRI (Fig. 4) and an abdominal CT (Fig. 5). While the brain MRI suggests three or four major “bumps” of the histogram at about values 25, 450, and 600, the abdominal CT doesn’t lend itself to such a description.

In the next step, we want to describe the histograms through analytical functions, to make them amenable for computational

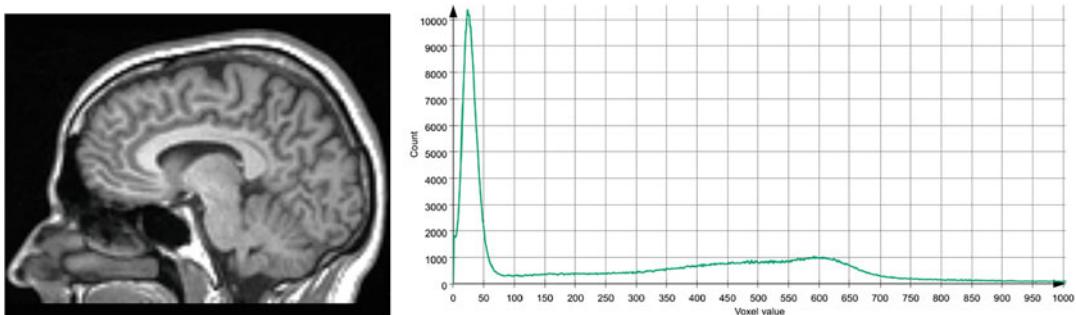


Fig. 4 Brain MRI (left) and histogram of gray values for one slice of a brain MRI

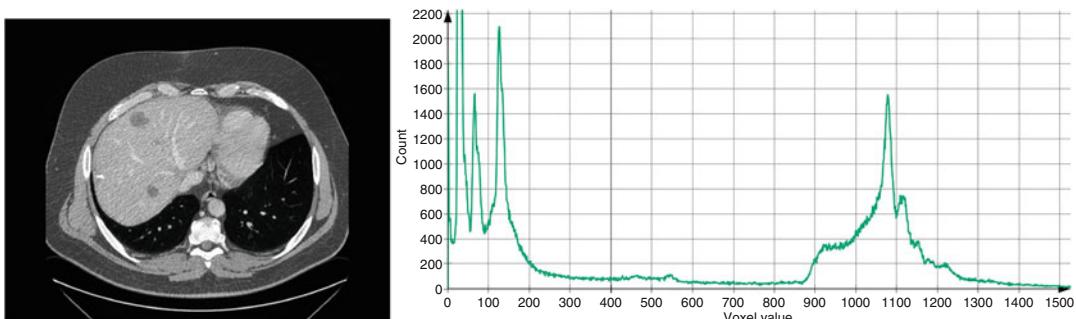


Fig. 5 Abdominal CT (left) and histogram of gray values for one slice of an abdominal CT

ends. This means we will aim to estimate an analytical description of the observations.

Expectation maximization (EM; *see Box 2*) is an algorithm suitable for this task. EM enables us to perform maximum likelihood estimation in the presence of unobserved (“latent”) variables and incomplete data—this being the default assumption when dealing with real data. Maximum likelihood estimation (MLE) is the process of finding parameters of a parametric distribution to most accurately match the distribution to the observations. In MLE, this is achieved by adapting the parameters steered by an error metric that indicates the closeness of the fit; in short, a parameter optimization algorithm.

Box 2: Expectation Maximization—Example

Focusing on our density estimate of the MRI data, we want to use expectation maximization (EM) to optimize the parameters of a fixed number of Gaussian functions adding up to the closest possible fit to the empirical shape of the histogram.

In our data, we observe “bumps” of the histogram. We can by image analysis determine that certain organs imaged by MRI lead to certain bumps in the histogram, since they are of different material and create different signal intensities. This, however, is unknown to EM—the so-called “latent” variables.

The EM algorithm has two parts, the expectation step and the maximization step. They can, with quite far-reaching omission of details, be sketched as follows:

Expectation takes each point (or a number of sampled points) of the distribution and *estimates the expectation* to which of the parameterized distribution to assign it to. Figuring out this assignment is the step of dealing with the “latent” variable of the observations.

Maximization iterates over all parameterized distributions and adjusts their parameters to match the assigned points as well as possible.

This process is iterated until a fitting error cannot be improved anymore.

A short introductory treatment of EM with examples and applications is presented in [7]. The standard reference for the algorithm is [8].

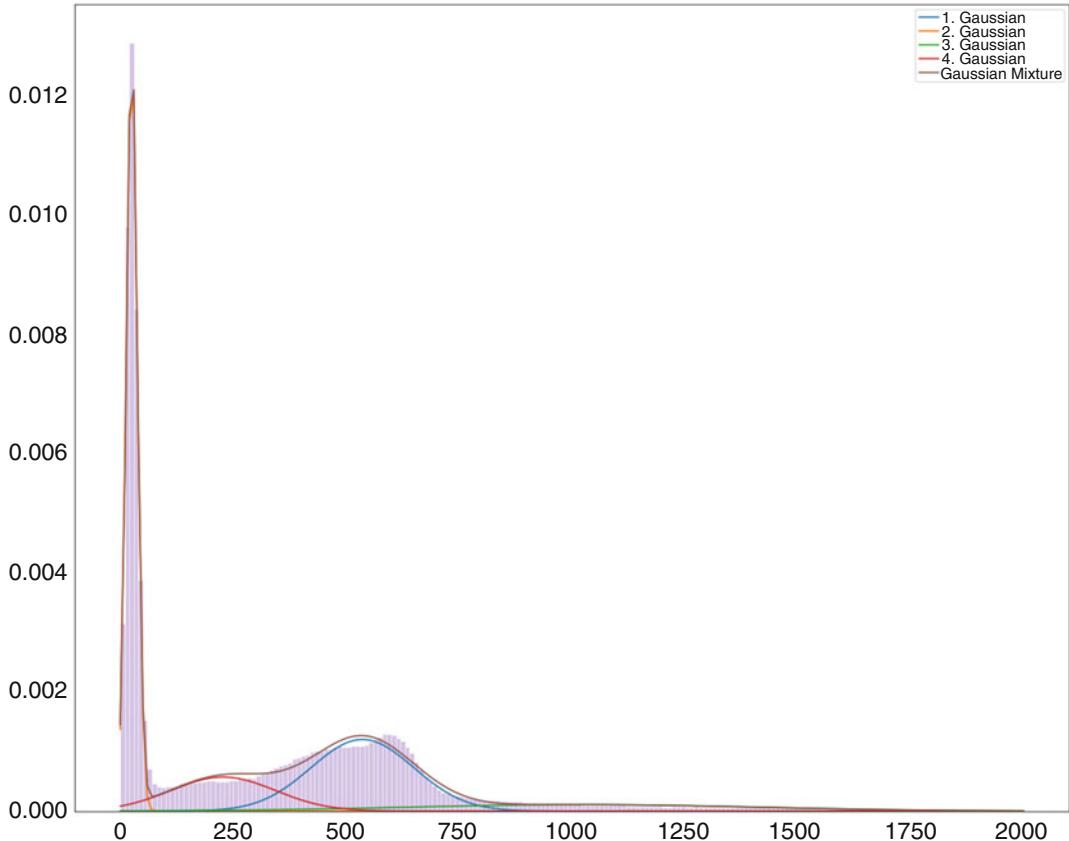


Fig. 6 A Gaussian mixture model (GMM) of four Gaussians was fit to the brain MRI data we have visualized as a histogram in Fig. 4

In Fig. 6, a mixture of four Gaussian distributions has been fit to the brain MRI voxel value data seen before.

It is tempting to model even more complex observations by mixing simple analytical distributions (e.g., Gaussian mixture models (GMMs)), but in general this will be intractable for two reasons. Firstly, realistic joint distributions will have an abundance of mixed maxima and therefore require a vast number of basic distributions to fit. Even basic normal distributions in high-dimensional parameter spaces are no longer functions with two parameters (μ, σ), but with a vector of means and a covariance matrix. Secondly, it is no longer trivial to sample from such high-dimensional joint distributions, and while some methods, among others Markov chain Monte Carlo methods, allow to sample from them, such numerical approaches are of such high computational complexity that it makes their use difficult in the context of deep neural network parameter estimation.

We will learn about alternatives. In principle, there are different approaches for density (distribution) estimation, direct distribution estimation, distribution approximation, or even more indirectly, by

using a simple surrogate distribution that is made to resemble the unknown distribution as good as possible through a mapping function. We will see this in the further elaboration of generative modeling approaches.

2.1.3 Estimators and the Expected Value

Assume we have found suitable mean values and standard deviations for three normal distributions that together approximate the shape of the MRI data density estimate to our satisfaction. Such a combination of normal (Gaussian) distributions is called a Gaussian mixture model (GMM), and sampling from such a GMM is straightforward. We are thus able to sample single pixels in any number, and over time we will sample them such that their density estimate or histogram will look similar to the one we started with.

However, if we want to generate a brain MRI image using a sampling process from our closed-form GMM representation of the distribution, we will notice that a very important notion wasn't respected in our approach. We start with one slice of 512×512 voxels and therefore randomly draw the required number of voxel values from the distribution. However, this will not yield an image that resembles one slice of a brain MRI, but will almost look like random noise, because we did not model the spatial relation of the gray values with respect to each other. Since the majority of voxels of a brain MRI are not independent of each other, drawing one new voxel from the distribution needs to depend on the spatial locations and gray values of all voxels drawn before. Neighboring voxels will have a higher likelihood of similar gray values than voxels far apart from each other, for example. More crucially, underneath the interdependence lies the image generation process: the image values observed in a real brain MRI stem from actual tissue—and this is what defines their interdependence. This means the anatomy of the brain indirectly reflects itself in the rules describing the dependency of gray values of one another.

For the modeling process, this implies that we cannot argue about single-voxel values and their likelihood, but we need to approach the generative process differently. One idea for a generative process has been implied in the above description already: pick a random location of the to-be-generated image and predict the gray value depending on all existing voxel values. Implemented with the method of mixture models, this results in unfathomably many distributions to be estimated, as for each possible “next voxel” location, any possible combination of already existing voxel numbers and positions needs to be considered. We will see in Subheading 5.1 on diffusion models how this general approach to image generation can still be made to work.

A different sequential approach to image generation has also been attempted, in which pixels are generated in a defined order, starting at the top left and scanning the image row by row across the columns. Again, the knowledge about the already produced

pixels is memorized and used to predict the next voxel. This has been dubbed the PixelRNN (Pixel Recurrent Neural Network), which lends its general idea from text processing networks [9].

Lastly, a direct approach to image generation could be formulated by representing or approximating the full joint distribution of all voxels in one distribution that is tangible and to sample all voxels *at once* from this. The full joint distribution in this approach remains implicit, and we use a surrogate. This will actually be the approach implemented in GANs, though not in a naive way.

Running the numbers of what a likelihood-based naive approach implies, the difficulties of making it work will become obvious. Consider an MRI image as the joint distribution of 512×512 voxels (one slice of our brain MRI), where we approximated the gray value distribution of one voxel with a GMM with six parameters. This results in a joint distribution of $512 \times 512 \times 6 = 1,572,864$ parameters. Conceptually, this representation therefore spans a 1,572,864-dimensional space, in which every one brain MRI slice will be one data point. Referring back to the histograms of CT and MRI images in the figures above, we have seen continuous lines with densities because we have collected all voxels of an entire medical image, which are many million. Still, we only covered one single dimension out of the roughly 1.5 million. Searching for the density in the 1,572,864-dimensional MRI-slice-space that is given by all collected brain MRI slices is the difficult task any generative algorithm has to solve. In this vastly large space, the brain MRI slices “live” in a very tiny region that is extremely hard to find. We say the images occupy a low-dimensional manifold within the high-dimensional space.

Consider the maximum likelihood formulation

$$\hat{\theta} = \arg \max_{\theta} \mathbb{E}_{x \sim P_{\text{data}}} \log Q_{\theta}(x|\theta) \quad (1)$$

where P_{data} is the unknown data distribution and Q_{θ} the distribution generated by the model which is parameterized by θ . θ can, for example, be the weights and biases of a deep neural network.¹ In other words, the result of maximum likelihood estimation is parameters $\hat{\theta}$ so that the product of two terms, out of which only the second depends on the choice of θ , is maximal. The first term is the expectation of x with regard to the real data distribution. The second term is the (log of) the conditional probability (likelihood) of seeing the example x given the choice of θ under the model Q_{θ} . Hence, maximizing the likelihood function means maximizing the probability that x is seen in Q_{θ} , which will be the case when Q matches P as closely as possible given the parametric form of Q .

¹ We will use θ when referring to parameters of models in general but designate parameters of neural networks with w in accordance with literature.

The maximum likelihood mechanism is very nicely illustrated in [10]. Here, it is also visually shown how finding the maximum likelihood estimate of parameters of the distribution can be done by working with partial derivatives of the likelihood function with respect to μ and σ^2 and seeking their extrema. The partial derivatives are called the score function and will make a reappearance when we discuss score-based and diffusion models later in Subheading 5.1 on advanced generative models.

2.1.4 Sampling from Distributions

When a distribution is a model of how observed values occur, then sampling from this distribution is the process of generating random new values that could have been observed, with a probability similar to the probability to observe this value in reality. There are two basic approaches to sampling from distributions: generating a random number from the uniform distribution (this is what a random number generator is always doing underneath) and feeding this number through the inverse cumulative density function (iCDF) of the distribution, which is the function that integrates the probability density function (PDF) of the distribution. This can only be achieved if the CDF is given in closed form. If it is not, the second approach to sampling can be used, which is called acceptance (or rejection) sampling. With f being the PDF, two random numbers x and y are drawn from the uniform distribution. The random x is accepted, if $f(x) > y$, and rejected otherwise.

Our use case, as we have seen, involves not only high-dimensional (multivariate) distributions but even more their joints, and they are not given in closed form. In such scenarios, sampling can be done still, using Markov chain Monte Carlo (MCMC) sampling, which is a framework using rejection sampling with added mechanisms to increase efficiency. While MCMC has favorable theoretic properties, it is still computationally very demanding for complex joint distributions, which leads to important difficulties in the context of sampling from distributions we are facing in the domain of image analysis and generation.

We are therefore at this point facing two problems: we can hardly hope to be able to estimate the density, and even if we could, we could practically not sample from it.

3 Generative Adversarial Networks

3.1 Generative vs. Discriminative Models

To emphasize the difficulty that generative models are facing, compare them to discriminative models. Discriminative models solve tasks like classification, detection, and segmentation, to name some of the most prominent examples. How classification models are in the class of discriminative models is obvious: discriminating examples is exactly classifying them. Detection models are also discriminative models, though in a broader sense, in that they classify the

detection proposals into accepted object detections or rejected proposals, and even the bounding box estimation, which is often solved through bounding box regression, typically involves the discriminative prediction of template boxes. Segmentation, on the other hand, for example, using a U-Net, is only the extension of classic discriminative approaches into a fast framework that avoids pixel-wise inference through the model. It is common to all these models that they yield output corresponding to their input, in the sense that they extract information from the input image (e.g., an organ segmentation, a classification, or even a textual description of the image content) or infer additional knowledge about it (e.g., a volume measurement or an assessment or prediction of a treatment success given the appearance of the image).

Generative models are fundamentally different, in that they generate output potentially without any concrete input, out of randomness. Still, they are supposed to generate output that conforms to certain criteria. In the most general form and intuitive formulation, their output should “look natural.” We want to further formalize the difference between the models in the following by using the perspective of distributions again. Figure 7 shows how discriminative and generative models have to construct differently complex boundaries in the representation space of the domain to accomplish their tasks.

Discriminative models take one example and map it to a label—e.g., the class. This is also true for segmentation models: they do this for each image voxel. The conceptual process is that the model has to estimate the probabilities that the example (or the voxel) comes from the distribution of the different available classes. The distributions of all possible appearances of objects of all classes do

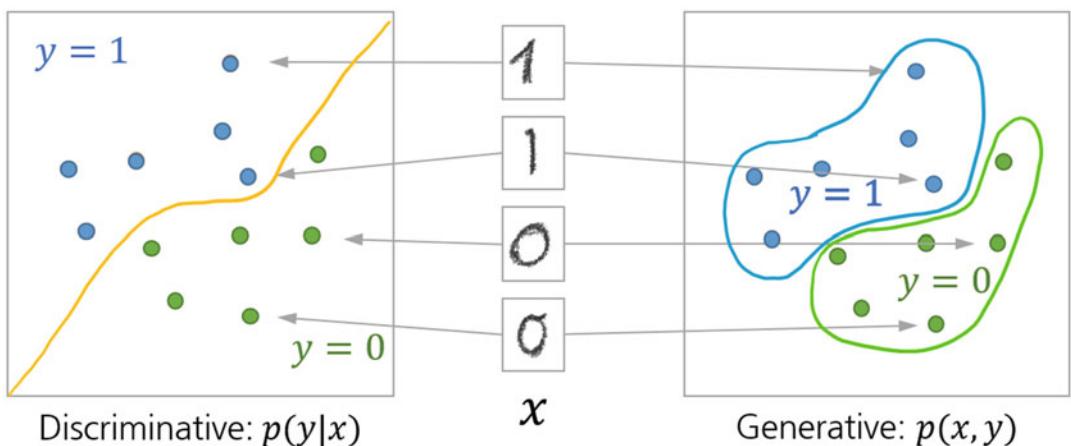


Fig. 7 The discriminative task compared to the generative task. Discriminative models only need to find the separating line between classes, while generative models need to delineate the part of space covering the classes (figure inspired by: <https://developers.google.com/machine-learning/gan/generative>)

not need to be modeled analytically for this to be successful. It is only important to know them locally—for example, it is sufficient to delineate their borders or overlaps with other distributions of other classes, but not all boundaries are important.

Generative models, on the other hand, are tasked to produce an example that is within a desired distribution. For this to work, the network has to learn the complete shape of this distribution. This is immensely complex, since all domains of practical importance in medical imaging are extremely high-dimensional and the distributions defining examples of interest within these domains are very small and hard to find. Also, they are neither analytically given nor normally distributed in their multidimensional space. But they have as many parameters as the output image of interest has voxels.

As already remarked, different other approaches were devised to generate output before GANs entered the scene. Among the trainable ones, approaches comprised (restricted) Boltzmann machines, deep belief networks, or generative stochastic networks, variational autoencoders, and others. Some of them involved feedback loops in the inference process (the prediction of a generated example) and were therefore unstable to train using backpropagation.

This was solved with the adversarial net framework proposed in 2014 by Goodfellow et al. [1]. They tried to solve the downsides like computational intractability or instability of such previous generative models by introducing the adversarial training framework.

To understand how GANs relate to one of the closest predecessors, the variational autoencoder, we will review their basic layout next. We will learn how elegantly the GAN paradigm turns the previously unsupervised approach to generative modeling into a supervised one, with the benefit of much more control over the training process.

3.2 Before GANs: Variational Autoencoders

Generative adversarial networks (GANs) haven't been the first or only attempt at generating realistically looking images (or any type of output, generally speaking). Apart from GANs, a related neural network-based approach to generative modeling is the variational autoencoder, which will be treated in more details below. Among other generative models with different approaches are as follows:

Flow-based models

This category of generative models attempt to model the data-generating distribution explicitly through an iterative process known as the normalizing flow [11], in which through repeated changes of variables a sequence of differentiable basis distributions is stacked to model the target distribution. The process is fully invertible, yielding models with desirable properties, since an

analytical solution to the data-generating distribution allows to directly estimate densities to predict the likelihood of future events, impute missing data points, and of course generate new samples. Flow-based models are computation-intensive. They can be categorized as a method that returns an explicit, tractable density. Another method in this category is, for example, the PixelRNN [9] or the PixelCNN [12] which also serves for conditional image generation. RealNVP [13] also uses a chain of invertible functions.

Boltzmann machines

work fundamentally differently. They also return explicit densities but this time only approximate the true target distribution. In this regard, they are similar to variational autoencoders, though their method is based on Markov chains, and not a variational approach. Deep Boltzmann machines have been proposed already in 2009, uniting a Markov chain-based loss component with a maximum likelihood-based component and showing good results on, at that time, highly complex datasets. [14] Boltzmann machines are very attractive but harder to train and use than other comparably powerful alternatives that exist today. This might change with future research, however.

Variational autoencoders (VAE) are a follow-up development of plain autoencoders, autoregressive models that in their essence try to reconstruct their input after transforming it, usually into a low-dimensional representation (*see* Fig. 8). This low-dimensional

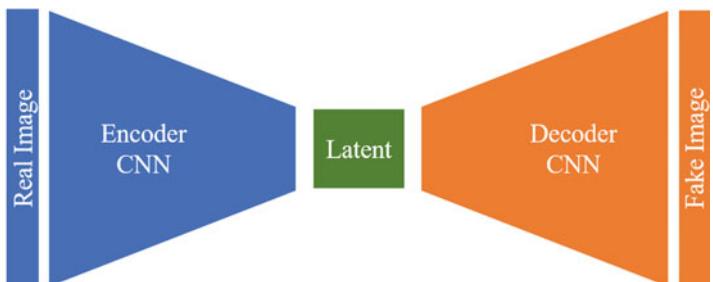


Fig. 8 Schematic of an autoencoder network. The encoder, for images, for example, a CNN with a number of convolutional and pooling layers, condenses the defining information of the input image into the variables of the latent space. The decoder, again convolutions, but this time with upsampling layers, recreates a representation in image space. Input and output images are compared in the loss function, which drives the gradient descent

representation is often termed the “latent space,” implying that here hidden traits of the data-generating process are coded, which are essential to the reconstruction process. This is very akin to the latent variables estimated by EM. In the autoencoder, the encoder will learn to code its input in terms of these latent variables, while the decoder will learn to represent them again in the source domain. In the following, we will be discussing the application to images though, in principle, both autoencoders and their variational variant are general mechanisms working for any domain.

We will later be interested in a behind-the-scene understanding of their modeling approach, which will be related to the employed loss function. We will then look at VAEs more extensively from the same vantage point: to understand their loss function—which is closest to the loss formulation of early GANs, the Kullback-Leibler divergence or KL divergence, D_{KL} .

With this tool in hand, we will examine how to optimize (train) a network with regard to KL divergence as the loss and understand key problems with this particular loss function. This will lead us to the motivation for a more powerful alternative.

3.2.1 From AE to VAE

VAEs are an interesting subject to study to emphasize the limits a loss function like KL divergence may place on a model. We will begin with a recourse to plain autoencoders to introduce the concept of learning a latent representation. We will then proceed to modify the autoencoder into a variational formulation which brings about the switch to a divergence measure as a loss function. From these grounds, we will then show how GANs again modified the loss function to succeed in high-quality image generation.

Figure 8 shows the schematic of a plain autoencoder (AE). As indicated in the sketch, input and output are of potentially very high dimensionality, like images. In between the encoder and decoder networks lies a “bottleneck” representation, which is, for example, a convolutional layer of orders of magnitude lower dimensionality (represented, for example, by a convolutional layer with only a few channels or a dense layer with a given low number of weights), which forces the network to find an encoding that preserves all information required for reconstruction.

A typical loss function to use when training the autoencoder is, for example, cross entropy, which is applicable for sigmoid activation functions, or simply the mean squared error (MSE). Any loss shall essentially force the AE to learn the identity function between input and output.

Let us introduce the notation for this. Let X be the input image tensor and X' the output image tensor. With f_w being the encoder function given as a neural network parameterized by weights and biases w and g_v the decoder function parameterized by v , the loss hence works to make $X = X' = g_v(f_w(X))$.

In a *variational* autoencoder,² things work differently. Auto-encoders like before use a fixed (deterministic) latent code to map the input to, while variational autoencoders will replace this with a distribution. We can call this distribution p_w , indicating the parameterization by w . It is crucial to understand that a choice was made here that imposes conditions on the latent code. It is meant to represent the input data in a variational way: in a way following Bayes' laws. Our mapping of the input image tensor X to the latent variable \mathbf{z} is by this choice defined by

- The prior probability $p_w(\mathbf{z})$
- The likelihood (conditional probability) $p_w(X|\mathbf{z})$
- The posterior probability $p_w(\mathbf{z}|X)$

Therefore, once we have obtained the correct parameters \hat{w} by training the VAE, we can produce a new output X' by sampling a $\mathbf{z}^{(i)}$ from the prior probability $p_{\hat{w}}(\mathbf{z})$ and then generate the example from the conditional probability through $X^{(i)} = p_{\hat{w}}(X|\mathbf{z}=\mathbf{z}^{(i)})$.

Obtaining the optimal parameters, however, isn't possible directly. The searched optimal parameters are those that maximize the probability that the generated example X' looks real. This probability can be rewritten as the aggregated conditional probabilities:

$$p_w(X^{(i)}) = \int p_w(X^{(i)}|\mathbf{z}) p_w(\mathbf{z}) d\mathbf{z}.$$

This, however, does not make the search any easier since we need to enumerate and sum up all \mathbf{z} . Therefore, an approximation is made through a surrogate distribution, parameterized by another set of parameters, q_v . Weng [15] shows in her explanation of the VAE the graphical model highlighting how q_v is a stand-in for the unknown searched p_w (see Fig. 9).

The reason to introduce this surrogate distribution actually comes from our wish to train neural networks for the decoding/encoding functions, and this requires us to back-propagate through the random variable, \mathbf{z} , which of course cannot be done. Instead, if we have control over the distribution, we can select it such that the reparameterization trick can be employed. We define q_v to be a multivariate Gaussian distribution with means and a covariance matrix that can be learned and a stochastic element multiplied to the covariance matrix for sampling [15, 16]. With this, we can back-propagate through the sampling process.

²Though variational autoencoders are in general not necessarily neural networks, in our context, we restrict ourselves to this implementation and stick to the notation with parameters w and v , where in many publications they are denoted θ and ϕ .

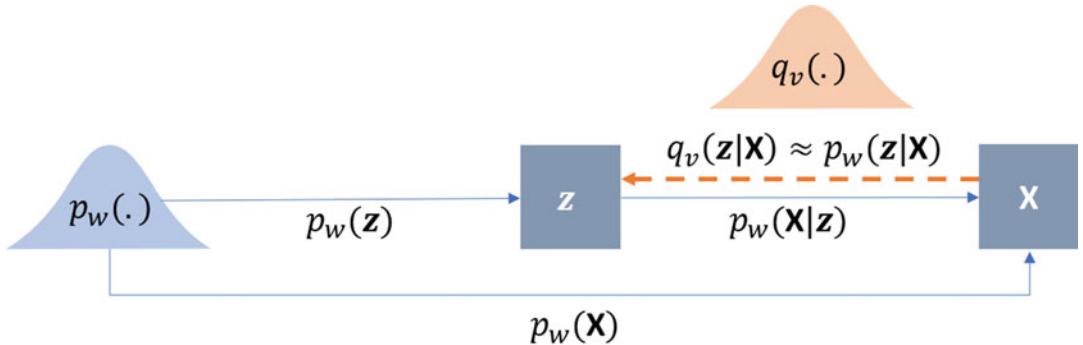


Fig. 9 The graphical model of the variational autoencoder. In a VAE, the *variational decoder* is $p_w(x|z)$, while the *variational encoder* is $q_v(z|x)$ (Figure after [15])

At this point, the two distributions need to be made to match: q_v should be as similar to p_w as possible. Measuring their similarity can be done in a variety of ways, of which Kulback-Leibler divergence (KL divergence or KLD) is one.

3.2.2 KL Divergence

A divergence can be thought of as an asymmetric distance function between two probability distributions, P and Q , measuring the similarity between them. It is a statistical distance which is not symmetric, which means it will not yield the same value if measured from P to Q or the other way around:

$$D_{\text{KL}}(P||Q) \neq D_{\text{KL}}(Q||P)$$

This can be seen when looking at the definition of KL divergence:

$$D_{\text{KL}}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \quad (2)$$

Sometimes, the measure D_{KL} is also called the relative entropy or information gain of P over Q , which also indicates the asymmetry.

To give the two distributions more meaning, let us associate them with a use case. P is usually the probability distribution of the example data, which can be our real images we wish to model, and is assumed to be unknown and high-dimensional. Q , on the other hand, is the modeled distribution, for example, parameterized by θ , similar to Eq. 1. Hence, Q is the distribution we can play with (in our case, optimize its parameters) to make them more similar to P . This means Q will get more informative with respect to the true P when we approach the optimal parameters.

Box 3: Example: Calculating D_{KL}

When comparing the two distributions given in Fig. 10, the calculation of the Kullback-Leibler divergence, D_{KL} , can explicitly be given by reading off the y values of the nine elements (columns) from Fig. 11 and inserting them into Eq. 2.

The result of this calculation is for

$$\begin{aligned} D_{\text{KL}}(P\|Q) &= \sum_x P(x) \log \frac{P(x)}{Q(x)} \\ &= 0.02 * \log \frac{.02}{.01} + 0.04 * \log \frac{.04}{.12} + \dots + 0.02 * \log \frac{.02}{.022} \\ &= 0.004 - 0.01 + \dots - 0.0002 \\ &= 0.0801 \end{aligned}$$

which we call “forward KL” as it calculates in the direction from the actual distribution P to the model distribution Q and for

$$\begin{aligned} D_{\text{KL}}(Q\|P) &= \sum_x Q(x) \log \frac{Q(x)}{P(x)} \\ &= 0.01 * \log \frac{0.01}{0.02} + 0.12 * \log \frac{0.12}{0.04} + \dots + 0.022 * \log \frac{0.022}{0.02} \\ &= -0.002 - 0.05 + \dots + 0.0002 \\ &= 0.0899 \end{aligned}$$

which we call “reverse KL.”

Note that in the example in Box 3, there is both a $P(X=x_i)$ and $Q(X=x_i)$ for each $i \in \{0, 1, \dots, 8\}$. This is crucial for KL divergence to work as a loss function.

3.2.3 Optimizing the KL Divergence

Examine what happens in forward and reverse KL if this condition is not satisfied for some i . If in forward KL P has values everywhere but Q has not (or extremely small values), the quotient in the log

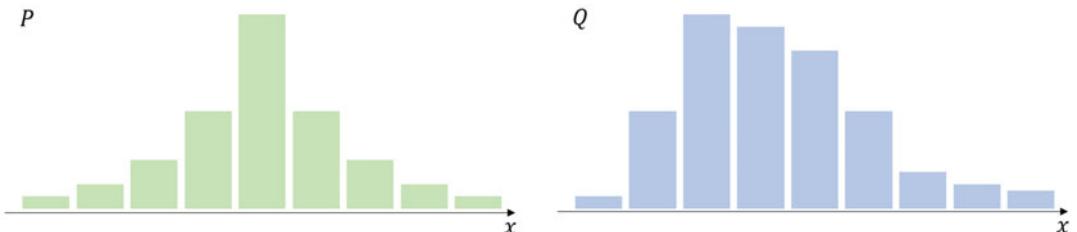


Fig. 10 Two distributions P and Q , here scaled to identical height

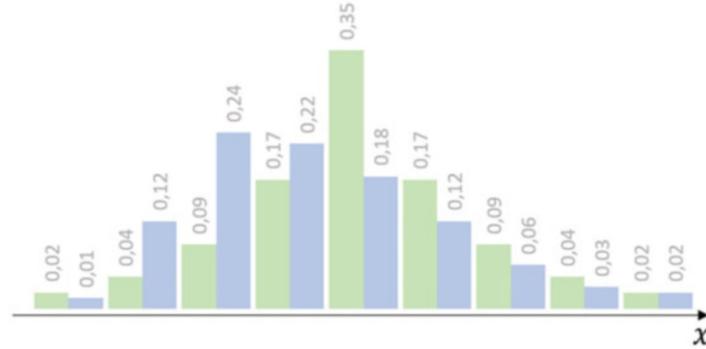


Fig. 11 The distributions P and Q , scaled to unit density, with added labels

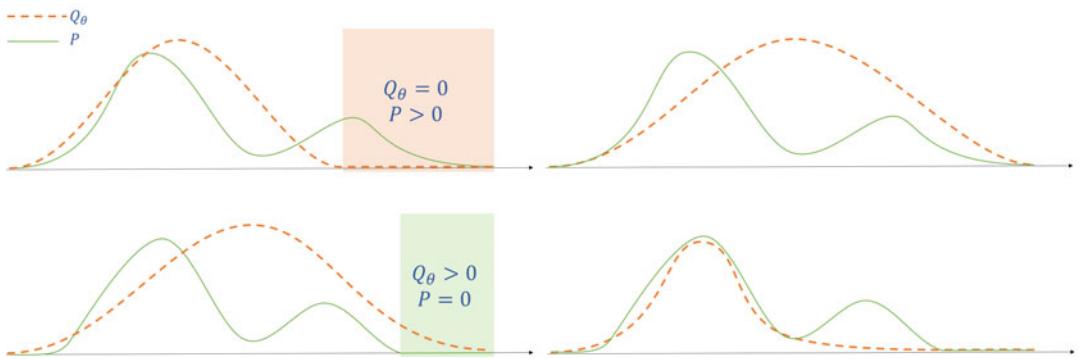


Fig. 12 The distributions P (solid) and Q_θ (dashed), in the initial configuration and after minimizing reverse KL $D_{KL}(Q_\theta|P)$. This time, in the initial configuration, Q_θ has values greater than 0 where P has not (marked with green shading)

function will tend to infinity by means of the division by almost zero, and the term will be very large.

In Fig. 12, we assume Q_θ to be a unimodal normal distribution, i.e., a Gaussian, while P is any empirical distribution. In the left plots of the figure, we show a situation before minimizing the forward/reverse KL divergence between P and Q_θ , in the right plots, the resulting shape of the Gaussian after minimization.

When in the minimization of forward KL $D_{KL}(P|Q_\theta)$ Q_θ is zero where P has values greater than zero, KL goes to infinity in these regions (marked area in the start configuration of the top row in Fig. 12), since the denominator in the log function goes to zero. This, in turn, drives the parameters of Q_θ to broaden the Gaussian to cover these areas, thereby removing the large loss contributions. This is known as the *mean-seeking* behavior of forward KL.

Conversely, in reverse KL (bottom row in Fig. 12), in the marked areas of the initial configuration, P is zero in regions where Q_θ has values greater than zero. This yields high-loss

contributions from the log denominator, in this case driving the Gaussian to remove these areas from Q_θ . Since we assumed a unimodal Gaussian Q , the minimization will focus on the largest mode of the unknown P . This is known as the *mode-seeking* behavior of reverse KL.

Forward KL tends to overestimate the target distribution, which is exaggerated in the right plot in Fig. 12. In contrast, reverse KL tends to underestimate the target distribution, for example, by dropping some of its modes. Since underestimation is the more desirable property in practical settings, reverse KL is the loss function of choice, for example, in variational autoencoders. The downside is that as soon as target distribution P and model distribution Q_θ have no overlap, KL divergence evaluates to infinity and is therefore uninformative. One countermeasure to take is to add noise to Q_θ , so that there is guaranteed overlap. This noise, however, is not desirable in the model distribution Q_θ since it disturbs the generated output.

Another way to remedy the problem of KL going to infinity is to adjust the calculation of the divergence, which is done in Jensen-Shannon divergence (JS divergence, D_{JS}) defined as

$$D_{JS} = \frac{1}{2}(D_{KL}(P\|M) + D_{KL}(Q_\theta\|M)), \quad (3)$$

where $M = \frac{P+Q_\theta}{2}$. In the case of nonoverlapping P and Q_θ , this evaluates to constant $\log 2$, which is still not providing information about the closeness but is computationally much friendlier and does not require the addition of a noise term to achieve numerical stability.

3.2.4 The Limits of VAE

In the VAE, reverse KL is used. Our optimization goal is maximizing the likelihood to produce realistic looking examples—ones with a high $p_w(x)$. Simultaneously, we want to minimize the difference between the real and estimated posterior distributions q_v and p_w . This can only be achieved through a reformulation of reverse KL [15]. After some rearranging of reverse KL, the loss of the variational autoencoder becomes

$$\begin{aligned} L_{VAE}(w, v) &= -\log p_w(X) + D_{KL}(q_v(\mathbf{z}|X)\|p_w(\mathbf{z}|X)) \\ &= -\mathbb{E}_{\mathbf{z} \sim q_v(\mathbf{z}|X)} \log p_w(X|\mathbf{z}) + D_{KL}(q_v(\mathbf{z}|X)\|p_w(\mathbf{z})) \end{aligned} \quad (4)$$

\hat{w} and \hat{v} are the parameters maximizing the loss.

We have seen how mode-seeking reverse KL divergence limits the generative capacity of variational autoencoders through the potential underrepresentation of all modes of the original distribution.

KL divergence and minimizing the ELBO also have a second fundamental downside: there is no way to find out how close our solution is to the obtainable optimum. We measure the similarity to the target distribution up to the KL divergence, but since the true $p_{\hat{w}}(\cdot)$ is unknown, the stopping criterion in the optimization has to be set by another metric, e.g., to a maximum number of iterations or corresponding to an improvement of the loss below some ε .

The original presentation of the variational autoencoder was given as one example of the general framework called the autoencoding variational Bayes. This publication presented the above ideas in a thorough mathematical formulation, starting from a directed graphical model that poses the abstract problem. The authors also develop the seminal “reparameterization trick” to make the loss formulation differentiable and with this to make the search for the autoencoder parameters amenable to gradient descent optimizers [16]. The details are beyond this introductory treatment.

3.3 The Fundamental GAN Approach

At the core of the adversarial training paradigm is the idea to create two players competing in a minimax game. In such games, both players have access to the same variables but have opposing goals, so that they will manipulate the variables in different directions.

Referring to Fig. 13, we can see the generative part in orange color, where random numbers are drawn from the latent space and, one by one, converted into a set of “fake images” by the generator

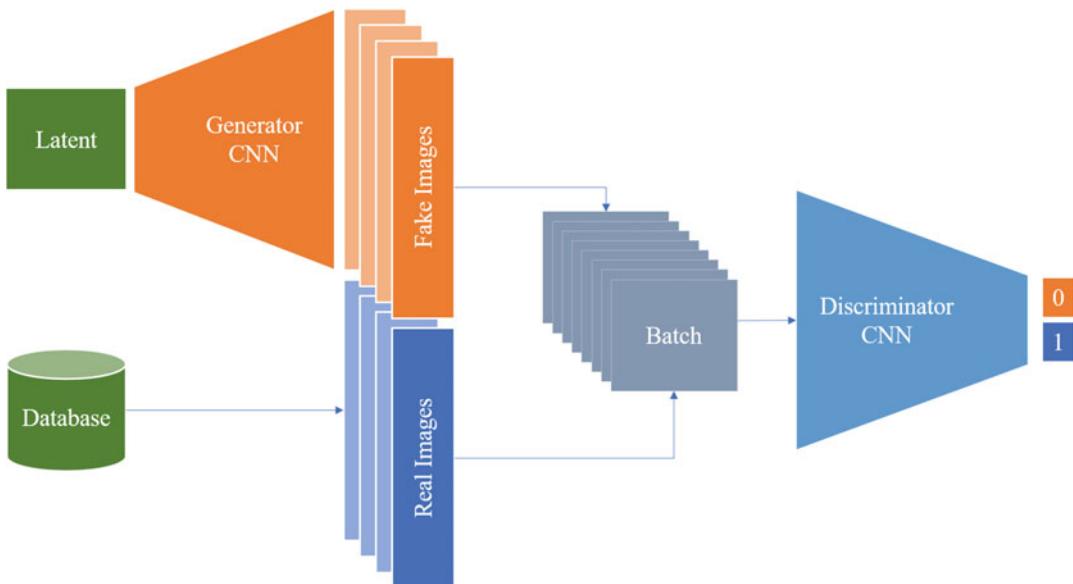


Fig. 13 Schematic of a GAN network. Generator (orange) creates fake images based on random numbers drawn from a latent space. These together with a random sample of real images are fed into the discriminator (blue, right). The discriminator looks at the batch of real/fake images and tries to assign the correct label (“0” for fake, “1” for real)

network, in the figure implemented by a CNN. Simultaneously, from a database of real images, a matching number of examples are randomly drawn. The real and fake images are composed into one batch of images which are fed into the discriminator. On the right side, the discriminator CNN is indicated in blue. It takes the batch of real and fake images and decides for each if it appears real (yielding a value close to “1”) or fake (“0”).

The error signal is computed from the number of correct assignments the discriminator can do on the batch of generated and real images. Both the generator and the discriminator can then update their parameters based on this same error signal. Crucially, the generator has the aim to *maximize* the error, since this signifies that it has successfully fooled the discriminator into taking the fake images for real, while the discriminator weights are updated to *minimize* the same error, indicating its success in telling true and fake examples apart. This is the core of the competitive game between generator and discriminator.

Let us introduce some abbreviations to designate GAN components. We will denote the generator and discriminator networks with G and D , respectively. The objective of GAN training is a game between generator and discriminator, where both affect a common loss function J , but in opposed directions. Formally, this can be written as

$$\min_G \max_D J(G, D),$$

with the GAN objective function

$$J(G, D) = \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_G} [1 - \log D(G(z))] \quad (5)$$

D will attempt to maximize J by maximizing the probability to assign the correct labels to real and generated examples: this is the case if $D(x) = 1$, maximizing the first loss component, and if $D(G(z)) = 0$, maximizing the second loss component. The generator G , instead, will attempt to generate realistic examples that the discriminator labels with “1,” which corresponds to a minimization of $\log(1 - D(G(z)))$.

3.4 Why Early GANs Were Hard to Train

GANs with this training objective implicitly use JS divergence for the loss, which can be seen by examining the GAN training objective. Consider the ideal discriminator D for a fixed generator. Its loss is minimal for the optimal discriminator given by [1]

$$\hat{D}(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_G(x)}. \quad (6)$$

Substituting \hat{D} in Eq. 5 yields (without proof) the implicit use of the Jensen-Shannon divergence if the above training objective is employed:

$$J(G, \hat{D}) = 2D_{\text{JS}}(p_{\text{data}} \| p_G) - \log 4. \quad (7)$$

This theoretical result shows that a minimum in the GAN training can be found when the Jensen-Shannon divergence is zero. This is achieved for identical probability distributions p_{data} and p_G or, equivalently, when the generator perfectly matches the data distribution [17].

Unfortunately, it also shows that this loss is, like KL divergence, only helpful when target distribution (i.e., data distribution) and model distribution have overlapping support. Therefore, added noise can be required to approximate the target distribution. In addition, the training criterion saturates if the discriminator in the early phase of training perfectly distinguishes between fake and real examples. The generator will therefore no longer obtain a helpful gradient to update its weights. An approach thought to prevent this was proposed by Goodfellow et al. [1]. The generator loss was turned from the minimization problem into a maximization problem that has the same fixed point in the overall minimax game but prevents saturation: instead of minimizing $\log(1 - D(G(z)))$, one maximizes $\log(D(G(z)))$ [1].

3.5 Improving GANs

GAN training has quickly become notorious for the difficulties it posed upon the researchers attempting to apply the mechanism to real-world problems. We have qualitatively attributed a part of these problems to the inherently difficult task of density estimation and motivated the intuition that while fewer samples might suffice to learn a decision boundary in a discriminative task, many more examples are required to build a powerful generative model.

In the following, some more light shall be shed on the reasons why GAN training might fail. Typical GAN problems comprise the following:

Mode dropping

is the phenomenon in forward KL caused by regions of the data distribution not being covered by the generator distribution, which implies large probabilities of samples coming from P_{data} and very small probabilities of originating from P_G . This drives forward KL toward infinity and punishes the generator for not covering the entire data distribution [18]. If all modes but one are dropped, one can call this mode collapse: the generator only generates examples from one mode of the distribution.

Poor convergence

can be caused by a discriminator learning to distinguish real and fake examples very early—which is also very likely to happen throughout the GAN training. This is rooted in the

observation that by the generative process that projects from a low-dimensional latent space into the high-dimensional p_G , the samples in p_G are not close to each other but rather inhabit “islands” [18]. The discriminator can learn to find them and thereby differentiate between true and false samples easily, which causes the gradients driving generator optimization to vanish [17].

Poor sample quality

despite a high log likelihood of the model is a consequence of the practical independence of sample quality and model log likelihood. Theis et al. [19] show that neither does a high log likelihood imply generated sample fidelity nor do visually pleasing samples imply a high log likelihood. Therefore, training a GAN with a loss function that effectively implements maximizing a log likelihood term is not an ideal choice—but exactly corresponds to KL minimization.

Unstable training

is a consequence of reformulating the generator loss into maximizing $\log D(G(z))$. It can be shown [18] that this choice effectively makes the generator struggle between a reverse KL divergence favoring mode-seeking behavior and a negative JS divergence actually driving the generator into examples different from the real data distribution.

There have been many subsequent authors touching these topics, but already Arjovsky and Bottou [18] have shown best practices of how to overcome these problems.

Among the solutions proposed for GAN improvements are some that prevent the generator from producing only too similar samples in one batch, some that keep the discriminator insecure about the true labels of real and fake examples, and more, which Creswell et al. [17] have summarized in their GAN overview. A collection of best practices compiled from these sources is presented in [Box 4](#). It is almost impossible to write a cookbook for successful, converging, stable GAN training. For almost every tip, there is a caveat or situation where it cannot be applied. The suggestions below therefore are to be taken with a grain of salt but have been used by many authors successfully.

Box 4: Best Practices for Stable GAN Training

General measures. GAN training is sensitive to hyperparameters, most importantly the learning rate. Mode collapse might already be mitigated by a lower learning rate. Also, different learning rates for generator and discriminator might help. Other typical measures are batch normalization (or instance normalization in case of small batch sizes; mind however that batch normalization can taint the randomness of latent vector sampling and in general should not be used in combination with certain GAN loss functions), use of transposed convolutions instead of parameter-free upsampling, and strided convolutions instead of down-sampling.

Feature matching. One typical observation is that neither discriminator nor generator converges. They play their “cat-and-mouse” game too effectively. The generator produces a good image, but the discriminator learns to figure it out, and the generator shifts to another good image, and so on.

A remedy for this is feature matching, where the ℓ_2 distance between the average feature vectors of real and fake examples is computed instead of a cross-entropy loss on the logits. Because per batch the feature vectors change slightly, this introduces randomness that helps to prevent discriminator overconfidence.

Minibatch discrimination. When the generator only produces very convincing but extremely similar images, this is an indication for mode collapse.

This can be counteracted by calculating a similarity metric between generated samples and penalizing the generator for too little variation. Minibatch discrimination is considered to be superior in performance to feature matching.

One-sided label smoothing. Deep classification models often suffer from overconfidence, focusing on only very few features to classify an image. If this happens in a GAN, the generator might figure this out and only produce the feature the discriminator uses to decide for a real example.

A simple measure to counteract this is to provide not a “1” as a label for the real images in the batch but a lower value. This way, the discriminator is penalized for overconfidence (when it returns a value close to “1”).

Cost function selection. Several sources list possible GAN cost functions. Randomly trying them one by one might work, but often some of the above measures, in particular learning rate and hyperparameter tuning, might be more successful first steps.

Besides these methods, one area of discussion concerned the question if there is a need of balancing discriminator and generator learning and convergence at all. The argument was that a converged discriminator will as well yield a training signal to the generator as a non-converged discriminator. Practically, however, many authors described carefully designed update schedules, e.g., updating the generator once per a given number of discriminator updates.

Many more ideas exist: weight updating in the generator using an exponential moving average of previous weights to avoid “forgetting,” different regularization and conditioning techniques, and injecting randomness into generator layers anew. Some we will encounter later, as they have proven to be useful in more recent GAN architectures.

Despite the recent advances in stabilizing GAN training, even the basic method described so far, with the improvements made in the seminal DCGAN publication [20], finds application until today, e.g., for the de novo generation of PET color images [21]. The usefulness of an approach as presented in their publication might be doubted, since the native PET data is obviously not colored. The authors use 2D histograms of the three-color channel combinations to compare true and fake examples. As we have discussed earlier, this is likely a poor metric since it does not allow insights into the high-dimension joint probability distribution underlying the data-generating process. Figure 14 shows an example comparison of some generated examples compared to original PET images.

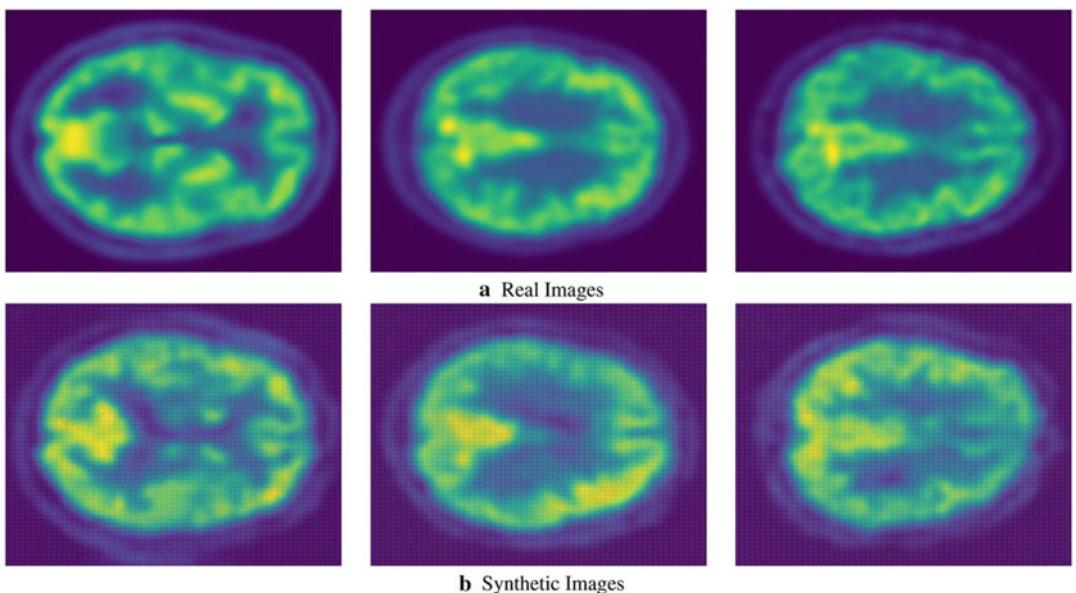


Fig. 14 PET images generated from random noise using a DCGAN architecture. Image taken from [21] (CC-BY4.0)

To address many of the GAN training dilemmas, Arjovsky and Bottou [18] have proposed to employ the Wasserstein distance as a replacement for KL or JS divergence already in their examination of the root causes of poor GAN training results and have later extended this into their widely anticipated approach we will focus on next [22, 23]. We will also see more involved and recent approaches to stabilize and speed up GAN training in later sections of this chapter (Subheading 4).

3.6 Wasserstein GANs

Wasserstein GANs were appealing to the deep learning and GAN scene very quickly after Arjovsky et al.’s [22] seminal publication because of a number of traits their inventors claimed they’d have. For one, Wasserstein GANs are based on the theoretical idea that the change of the loss function to the Wasserstein distance should lead to improved results. This combined with the reported benchmark performance would already justify attention. But Wasserstein GANs additionally were reported to train much more stably, because, as opposed to previous GANs, the discriminator would be trained to convergence in every iteration, instead of demanding a carefully and heuristically found update schedule for generator and discriminator. In addition, the loss was directly reported to correlate with visual quality of generated results, instead of being essentially meaningless in a minimax game.

Wasserstein GANs are therefore worth an in-depth treatment in the following sections.

3.6.1 The Wasserstein (Earthmover) Distance

The Wasserstein distance figuratively measures how, with an optimal transport plan, mass can be moved from one configuration to another configuration with minimal work. Think, for example, of heaps of earth. Figure 15 shows two heaps of earth, P and Q (discrete probability distributions), both containing the same amount of earth in total, but in different concrete states x and y out of all possible states.

Work is defined as the shovelfuls of earth times the distance it is moved. In the three rows of the figure, earth is moved (only within one of P or Q , not from one to the other), in order to make the configuration identical. First, one shovelful of earth is moved one pile further, which adds one to the Wasserstein distance. Then, two shovelfuls are moved three piles, adding six to the final Wasserstein distance of $D_W = 7$.

Note that in an alternative plan, it would have been possible to move two shovelfuls of earth from p_4 to p_1 (costing six) and one from p_4 to p_3 , which is the inverse transport plan of the above, executed on P , and leading to the same Wasserstein distance. The Wasserstein distance is in fact a distance, not a divergence, because it yields the same result regardless of the direction. Also note that

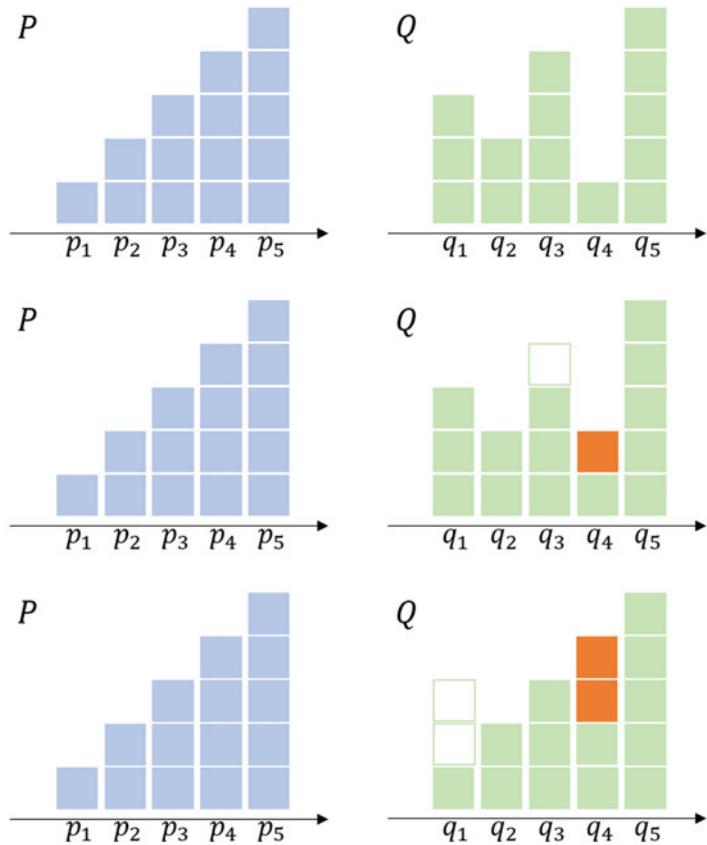


Fig. 15 One square is one shovel full of earth. Transporting the earth shovel-wise from pile to pile amasses performed work: the Wasserstein (earthmover) distance. The example shows a Wasserstein distance of $D_W = 7$

we implicitly assumed that P and Q share their support,³ but that in case of disjunct support, only a constant term would have to be added, which grows with the distance between the support regions.

Many other transport plans are possible, and others can be equally cheap (or even cheaper—it is left to the reader to try this out). Transport plans need not modify only one of the stocks but can modify both to reach the optimal strategy to make them identical. Algorithmically, the optimal solution to the question of the optimal transport plan can be found by formulating it as a linear programming problem. However, enumerating all transport plans and computing the linear programming algorithm are intractable for larger and more complex “heaps of earth.” Any nontrivial GAN will need to estimate transport of such complex “heaps,” so they

³The support, graphically, is the region where the distribution is not equal to zero.

suffer this intractability problem. Consequently, in practice, a different approach must be taken, which we will sketch below.⁴

Formalizing the search for the optimal transport plan, we look at all possible joint distributions of our P and Q , forming the set of all possible transport plans, and denote this set $\Pi(P, Q)$, implying that for all $\gamma \in \Pi(P, Q)$, P and Q will be their marginal distributions.⁵ This, in turn, means that by definition $\sum_x \gamma(x, y) = P(y)$ and $\sum_y \gamma(x, y) = Q(x)$.

For one concrete transport plan γ that works between a state x in P and a state y in Q we are interested in the optimal transport plan $\gamma(x, y)$. Let $\|x - y\|$ be the Euclidian distance to shift earth between x and y , and then multiplying this with every value of γ (the amount of earth shifted) leads to

$$D_W(P, Q) = \inf_{\gamma \in \Pi} \sum_{x, y} \|x - y\| \gamma(x, y),$$

which can be rewritten to obtain

$$D_W(P, Q) = \inf_{\gamma \sim \Pi(P, Q)} \mathbb{E}_{(x, y) \sim \gamma} \|x - y\|. \quad (8)$$

It measures both the distance of two distributions with disjunct support and the difference between distributions with perfectly overlapping support because it includes both, the shifting of earth and the distance to move it.

Practically, though, this result cannot be used directly, since the Linear Programming problem scales exponentially with the number of dimensions of the domain of P and Q , which are high for images. To our disadvantage, we additionally need to differentiate the distance function if we want to use it for deep neural network training using backpropagation. However, we cannot obtain a derivative from our distance function in the given form, since, in the linear programming (LP) formulation, our optimized distribution (as well as the target distribution) end up as constraints, not parameters.

Fortunately, we are not interested in the transport plan γ itself, but only in the distance (of the optimal transport plan). We can therefore use the dual form of the LP problem, in which the constraints of the primal form become parameters. With some clever definitions, the problem can be cast into the dual form, finally yielding

⁴ An extensive treatment of Wasserstein distance and optimal transport in general is given in the 1.000-page treatment of Villani's book [24], which is freely available for download.

⁵ This section owes to the excellent blog post of Vincent Herrmann, at <https://vincentherrmann.github.io/blog/wasserstein/>. Also recommended is the treatment of the "Wasserstein GAN" paper by Alex Irpan at <https://www.alexirpan.com/2017/02/22/wasserstein-gan.html>. An introductory treatment of Wasserstein distance is also found in [25, 26].

$$D_W(P, Q) = \|f\|_{L \leq 1} \sup \mathbb{E}_{x \sim P} f(x) - \mathbb{E}_{x \sim Q} f(x)$$

with a function f that has to adhere to a constraint called the 1-Lipschitz continuity constraint, which requires f to have a slope of at most magnitude 1 everywhere. f is the neural network, and more specifically for a GAN, the discriminator network. 1-Lipschitzness can be achieved trivially by clipping the weights to a very small interval around 0.

3.6.2 Implementing WGANs

To implement the distance as a loss function, we rewrite the last result again as

$$D_W(P, Q) = \max_{\pi \in \mathcal{W}} \mathbb{E}_{x \sim P} [D_\pi(x)] - \mathbb{E}_{z \sim Q} [D_\pi(G_\pi(z))]. \quad (9)$$

Note that in opposition to other GAN losses we have seen before, there is no logarithm anymore, because, this time, the “discriminator” is no longer a classification network that should learn to discriminate true and fake samples but rather serves as a “blank” helper function that during training learns to estimate the Wasserstein distance between the sets of true and fake samples.

Box 5: Spectral Normalization

Spectral normalization is applied to the weight matrices of a neural network to ensure a boundedness of the error function (e.g., Lipschitzness of the discriminator network in the WGAN context). This helps convergence like any other normalization method, as it provides a guaranty that gradient directions are stable around the current point, allowing larger step widths.

The **spectral norm** (or matrix norm) measures how far a matrix A can stretch a vector x :

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

The numerical value of the spectral norm of A can be shown to be just its maximum singular value. To compute the maximum singular value, an algorithmic idea helps: the power iteration method, which yields the maximal eigenvector.

Power iteration uses the fact that any matrix will rotate a random vector toward its largest eigenvector. Therefore, by iteratively calculating $\frac{Ax}{\|Ax\|}$, the largest eigenvector is obtained eventually.

In practice, it is observed that a single iteration is already sufficient to achieve the desired normalizing behavior.

Consequently, the key ingredient is the Lipschitzness constraint of the discriminator network,⁶ and how to enforce this in a stable and regularized way. It soon turned out that weight clipping is not an ideal choice. Rather, two other methods have been proposed: the gradient penalty approach and normalizing the weights with the spectral norm of the weight matrices.

Both have been added to the standard catalogue of performance-boosting measures in GAN training ever since, where in particular spectral normalization (cf. [Box 5](#)) is attractive as it can be implemented very efficiently, has a sound theoretical and mathematical foundation, and ensures stable and efficient training.

3.6.3 Example Application: Brain Abnormality Detection Using WGAN

One of the first applications of Wasserstein GANs in a practical use case was presented in the medical domain, specifically in the context of attributing visible changes of a diseased patient with respect to a normal control to locations in the images [27]. The way this detection problem was cast into a GAN approach (and then solved with a Wasserstein GAN) was to delineate the regions that make the images of a diseased patient look “diseased,” i.e., find the residual region, that, if subtracted from the diseased-looking image, would make it look “normal.”

[Figure 16](#) shows the construction of the VA-GAN architecture with images from a mocked dataset for illustration. For the authors’ results, see their publication and code repository.⁷

For their implementation, the authors note that neither batch normalization nor layer normalization helped convergence and hypothesize that the difference between real and generated examples may be a reason that in particular batch normalization may in fact have an adverse effect especially during the early training phase. Instead, they impose an ℓ_1 norm loss component on the U-Net-generated “visual (feature) attribution” (VA) map to ensure it to be a minimal change to the subject. This serves to prevent the generator from changing the subject into some “average normal” image that it may otherwise learn. They employ an update regime that trains the critic network for more iterations than the generator, but doesn’t train it to convergence as proposed in the original WGAN publications. Apart from these measures, in their code repository, the authors give several practical hints and heuristics that may stabilize the training, e.g., using a *tanh* activation for the generator or exploring other dropout settings and in general using a large enough dataset. They also point out that the Wasserstein distance isn’t suited for model selection since it is too unstable and not directly correlated to the actual usefulness of the trained model.

⁶The discriminator network in the context of continuous generator loss functions like the Wasserstein-based loss is called a “critique” network, as it no longer discriminates but yields a metric. For ease of reading, this chapter sticks to the term “discriminator.”

⁷<https://github.com/baumgach/vagan-code>.

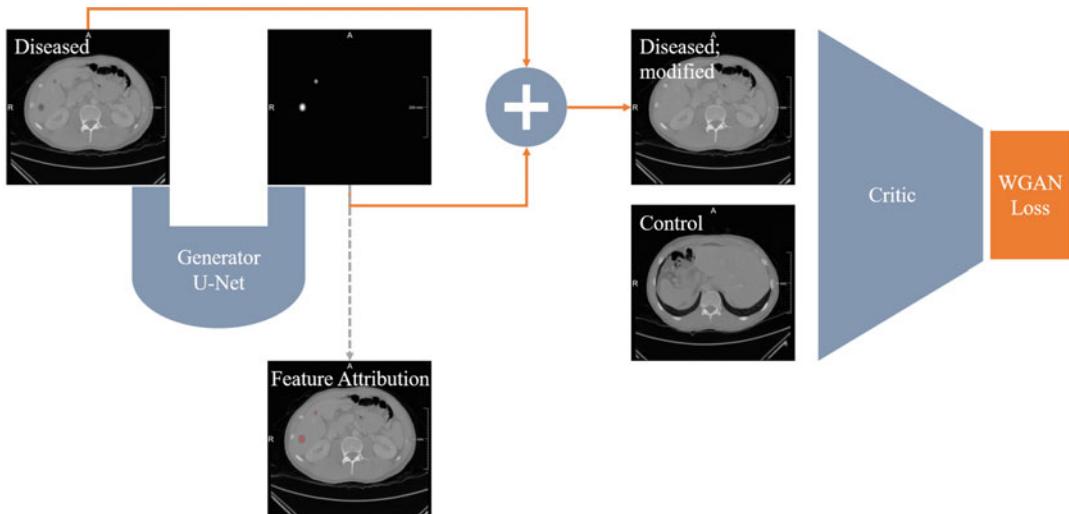


Fig. 16 An image of a diseased patient is run through a U-Net with the goal to yield a map that, if added to the input image, results in a modified image that fools the discriminator (“critique”) network into classifying it as a “normal” control. The map can be interpreted as the regions attributed to appear abnormal, giving rise to the name of the architecture: visual attribution GAN (VA-GAN)

This is one more reason to turn in the next section to an important topic in the context of validation for generative models: How to quantify their results?

3.7 GAN Performance Metrics

One imminent question has so far been postponed, though it implicitly plays a crucial role in the quest for “better” GANs: How to actually measure the success of a GAN or the performance in terms of result quality?

GANs can be adapted to solve image analysis tasks like segmentation or detection (cf. Subheading 3.6.3). In such cases, the quality and success can be measured in terms of task-related performance (Jaccard/Dice coefficient for segmentation, overlap metrics for detection etc.).

Performance assessment is less trivial if the GAN is meant to generate unseen images from random vectors. In such scenarios, the intuitive criterion is how convincing the generated results are. But convincing to whom? One could expose human observers to the real and fake images, ask them to tell them apart, and call a GAN better than a competing GAN if it fools the observer more consistently.⁸ Since this is practically infeasible, metrics were sought that provide a more objective assessment.

⁸ In fact, there is only very little research on the actual performance of GANs in fooling human observers, though guides exist on how to spot “typical” GAN artifacts in generated images. These are older than the latest GAN models, and it can be hypothesized that the lack of such literature is indirect confirmation of the overwhelming capacity of GANs to fool human observers.

The most widely used way to assess GAN image quality is the Fréchet inception distance (FID). This distance is conceptually related to the Wasserstein distance. It has an analytical solution to calculate the distance of Gaussian (normal) distributions. In the multivariate case, the Fréchet distance between two distributions X and Y is given by the squared distance of their means μ_X (resp. μ_Y) and a term depending on the covariance matrix describing their variances Σ_X (resp. Σ_Y):

$$d(X, Y) = \|\mu_X - \mu_Y\|^2 + \text{Tr}(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y}). \quad (10)$$

The way this distance function is being used is often the score, which is computed as follows:

- Take two batches of images (real/fake, respectively).
- Run them through a feature extraction or embedding model. For FID, the inception model is used, pretrained on ImageNet. Retain the embeddings for all examples.
- Fit each one multivariate normal distribution to the embedded real/fake examples.
- Calculate their Fréchet distance according to the analytical formula in Eq. 10.

This metric has a number of downsides. Typically, if computed for a larger batch of images, it decreases, although the same model is being evaluated. This bias can be remedied, but FID remains the most used metric still. Also, if the inception network cannot capture the features of the data FID should be used on, it might simply be uninformative. This is obviously a grave concern in the medical domain where imaging features look much different from natural images (although, on the other hand, transfer learning for medical classification problems proved to work surprisingly well, so that apparently convolutional filters trained on photographs also extract applicable features from medical images). In any case, the selection of the pretrained embedding model brings a bias into the validation results. Lastly, the assumption of a multivariate normal distribution for the inception features might not be accurate, and only describing it through their means and covariances is a severe reduction of information. Therefore, a qualitative evaluation is still required.

One obvious additional question arises: If the ultimate metric to judge the quality of the generator is given by, for example, the FID, why can't it be used as the optimization goal instead of minimizing a discriminator loss? In particular, as the Fréchet distance is a variant of the Wasserstein distance, an answer to this question is not obvious. In fact, feature matching as described in [Box 4](#) exactly uses this type of idea, and likewise, it has been partially adopted in recent GAN architectures to enhance the stability of training with a more fine-grained loss component than a pure categorical cross-entropy loss on the “real/fake” classification of the discriminator.

Related recent research is concerned with the question how generated results can automatically be detected to counteract fraudulent authors. So-called forensic algorithms detect patterns that point out generated images. This research puts up the question how to detect fake images reliably. Solutions based on different analysis directions encompass image fingerprinting and frequency-domain analysis [28–31].

4 Selected GAN Architectures You Should Know

In the following, we will examine some GAN architectures and GAN developments that were taken up by the medical community or that address specific needs that might make them appealing, e.g., for limited data scenarios.

4.1 Conditional GAN

GANs cannot be told what to produce—at least that was the case with early implementations. It was obvious, though, that a properly trained GAN would imprint the semantics of the domain onto its latent space, which was evidenced by experiments in which the latent space was traversed and images of certain characteristics could be produced by sampling accordingly. Also, it was found that certain dimensions of the latent space can correspond to certain features of the images, like hair color or glasses, so that modifying them alone can add or take away such visible traits.

With the improved development of conditional GANs [32] following a number of GANs that modeled the conditioning input more explicitly, another approach was introduced that was based on the U-Net architecture as a generator and a favorable discriminator network that values local style over a full-image assessment.

Technically, the formulation of a conditional GAN is straightforward. Recalling the value function (learning objective) of GANs from Eq. 5,

$$J(G, D) = \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_G} [1 - \log D(G(z))],$$

We now want to condition the generation on some additional knowledge or input. Consequently, both the generator G and the discriminator D will receive an additional “conditioning” input, which we call x . This can be a class label but also any other associated information. Very commonly, the additional input will be an image, as, for example, for image translation application (e.g., transforming from one image modality to another such as, for instance, MRI to CT). The result is the cGAN objective function:

$$J_{\text{cGAN}}(G, D) = \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x|y)] + \mathbb{E}_{z \sim p_G} [1 - \log D(G(z|y))] \quad (11)$$

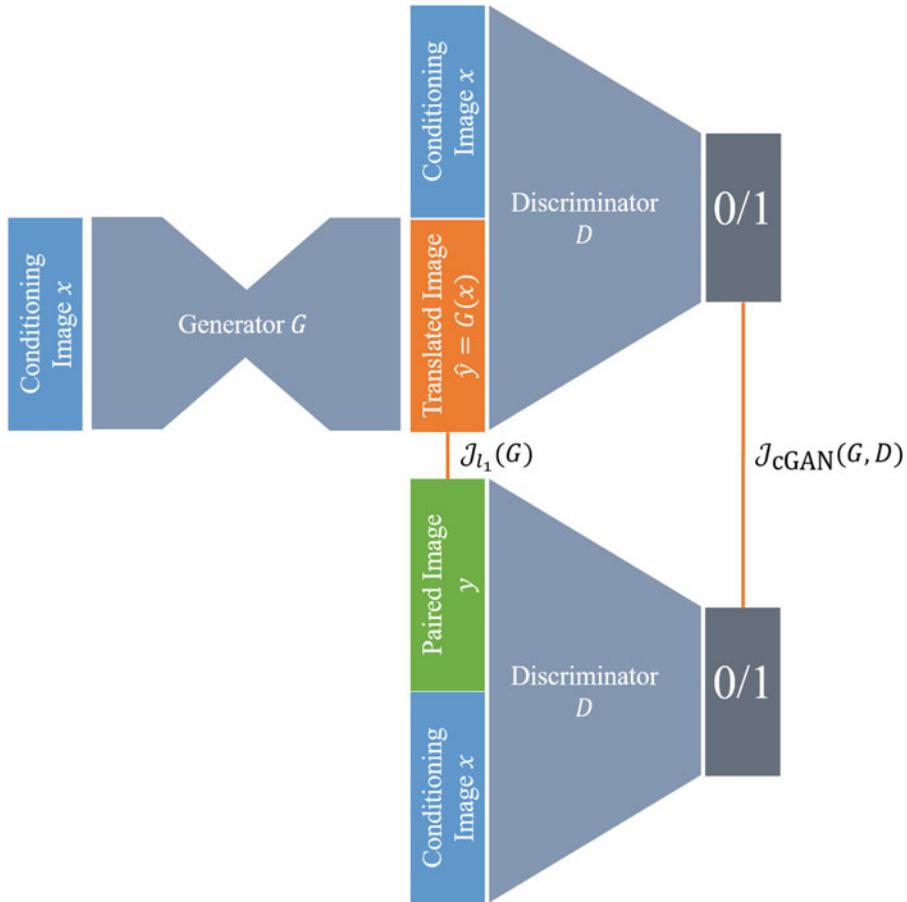


Fig. 17 A possible architecture for a cGAN. Left: the generator network takes the base images x as input and generates a translated image \hat{y} . The discriminator receives either this pair of images or a true pair x, y (right). The additional generator reconstruction loss (often a ℓ_1 loss) is calculated between y and \hat{y}

Isola et al. [32] describe experiments with MNIST handwritten digits, where a simple generator with two layers of fully connected neurons was used, and similarly for the discriminator. x was set to be the class label. In a second experiment, a CNN creates a feature representation of images, and the generator is trained to generate textual labels (choosing from a vocabulary of about 250.000 encoded terms) for the images conditioned on this feature representation.

Figure 17 shows a possible architecture to employ a cGAN architecture for image-to-image translation. In this diagram, the conditioning input is the target image that the trained network shall be able to produce based on some image input. The generator network therefore is a U-Net. The discriminator network can be implemented, for example, by a classification network. This network always receives two inputs: the conditioning image (x in Fig. 17) and either the generated output \hat{y} or the true paired image y .

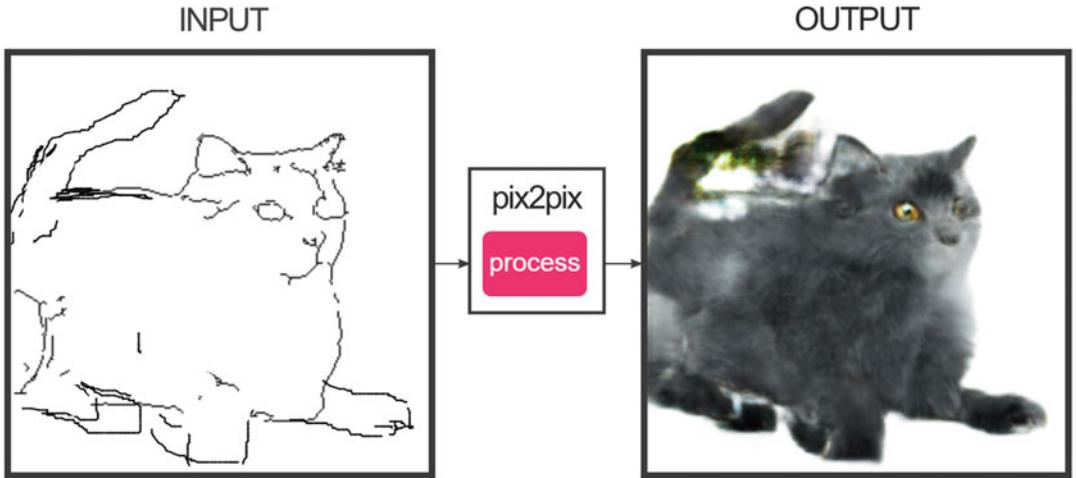


Fig. 18 Input and output of a pix2pix experiment. Online demo at <https://affinelayer.com/pixsrv/>

Note that the work of Isola et al. [32] introduces an additional loss term on the generator that measures the ℓ_1 distance between the generated and ground truth image, which is (with variables as in Eq. 11)

$$\mathcal{I}_{\ell_1}(G) = \mathbb{E}_{x,y,z} \|y - G(x, z)\|_1,$$

where $\|\cdot\|_1$ is the ℓ_1 norm.

The authors do not further justify this loss term apart from stating that ℓ_1 is preferred over ℓ_2 to encourage less blurry results. It can be expected that this loss component provides a good training signal to the generator when the discriminator loss doesn't, e.g., in the beginning of the training with little or no overlap of target and parameterized distributions. The authors propose to give the ℓ_1 loss orders of magnitudes more weight than the discriminator loss component to value accurate translations of images over "just" very plausible images in the target domain.

The cGAN, namely, in the configuration with a U-Net serving as the generative network, was very quickly adopted by artists and scientists, thanks to the free implementation pix2pix.⁹ One example created with pix2pix is given in Fig. 18, where the cGAN was trained to produce cat images from line drawings.

One application in the medical domain was proposed, for example, by Senaras et al. [33]. The authors used a U-Net as a generator to produce a stained histopathology image from a label image that has two distinct labels for two kinds of cell nuclei. Here, the label image is the conditioning input to the network. Consequently, the discriminator network, a classification CNN tailored to

⁹ <https://github.com/phillipi/pix2pix>.

the patch-based classification of slides, receives two inputs: the histopathology image and a label image.

Another example employed an augmented version of the conditional GAN to translate CT to MR images of the brain, including a localized uncertainty estimate about the image translation success. In this work, a Bayesian approach to model the uncertainty was taken by including dropout layers in the generator model [34].

Lastly, a 3D version of the pix2pix approach with a 3D U-Net as a generative network was devised to segment gliomas in multi-modal brain MRI using data from the 2020 International Multi-modal Brain Tumor Segmentation (BraTS) challenge [35]). The authors called their derived model vox2vox, alluding to the extension to 3D data [36].

More conditioning methods have been developed over the years, some of which will be sketched further on. It is common to this type of GANs that paired images are required to train the network.

4.2 CycleGAN

While cGANs require paired data for the gold standard and conditioning input, this is often hard to come by, in particular in medical use cases. Therefore, the development of the CycleGAN set a milestone as it alleviates this requirement and allows to train image-to-image translation networks without paired input samples.

The basic idea in this architecture is to train two mapping functions between two domains and to execute them in sequence so that the resulting output is considered to be in the origin domain again. The output is compared against the original input, and their ℓ_1 or ℓ_2 distance establishes a novel addition to the otherwise usual adversarial GAN loss. This might conceptually remind one of the autoencoder objectives: reproduce the input signal after encoding and decoding; only this time, there is no bottleneck but another interpretable image space. This can be exploited to stabilize the training, since the sequential concatenation of image translation functions, which we will call G and F , can be reversed. Figure 19 shows a schematic of the overall process (left) and one incarnation of the cycle, here from image domain X to Y and back (middle).

CycleGANs employ several loss terms in training: two adversarial losses $\mathcal{J}(G, D_Y)$ and $\mathcal{J}(F, D_X)$ and two cycle consistency losses, of which one $\mathcal{J}_{\text{cyc}}(G, F)$ is indicated rightmost in Fig. 19. Zhu et al. [37] presented the initial publication with a participation of the cGAN author Isola [37]. The cycle consistency losses are ℓ_1 losses in their implementation, and the GAN losses are least square losses instead of negative log likelihood, since more stable training was observed with this choice.

Almahairi et al. [38] provided an augmented version [38], noting that the original implementation suffers from the inability to generate stochastic results in the target domain Y but rather learns a one-to-one mapping between X and Y and vice versa. To

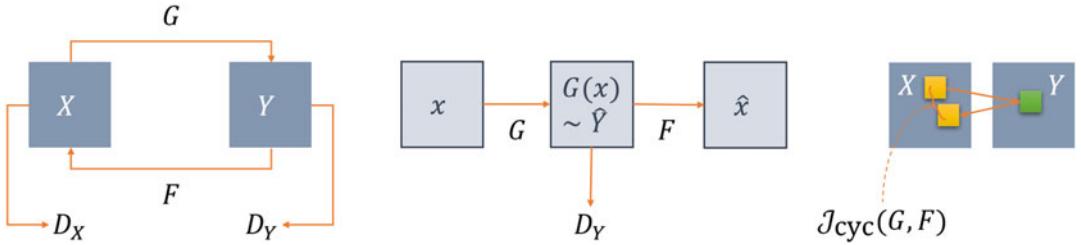


Fig. 19 Cycle GAN. Left: image translation functions G and F convert between two domains. Discriminators D_X and D_Y give adversarial losses in both domains. Middle: for one concrete translation of an image x , the translation to Y and back to X is depicted. Right: after the translation cycle, the original and back-translated result are compared in the *cycle consistency loss*

alleviate this problem, the generators are conditioned on one latent space each for both directions, so that, for the same input $x \in X$, G will now produce multiple generated outputs in Y depending on the sample from the auxiliary latent space (and similarly in reverse). Still, F has to recreate a \hat{x} minimizing the cycle consistency loss for each of these samples. This also remedies a second criticism brought forward against vanilla CycleGANs: these networks can learn to hide information in the (intermediate) target image domain that fool the discriminator but help the backward generator to minimize the cycle consistency loss more efficiently [39]. Chu et al. [39] use adaptive histogram equalization to show that in visually empty regions of the intermediate images information is present. This is a finding reminiscent of adversarial attacks, which the authors elaborate on in their publication.

Zhang et al. [40] show a medical application. In their work, a CycleGAN has been used to train image translation and segmentation models on unpaired images of the heart, acquired with MRI and CT and with gold standard expert segmentations available for both imaging datasets. The authors proposed to learn more powerful segmentation models by enriching both datasets with artificially generated data. To this end, MRIs are converted into CT contrast images and vice versa using GANs. Segmentation models for MRI and CT are then trained on dataset consisting of original images and their expert segmentations and augmented by the converted images, for which expert segmentations can be carried over from their original domain. To achieve this, it is of importance that the converted (translated) images accurately depict the shape of the organs as expected in the target domain, which is enforced using the shape consistency loss.

In the extended setup of the CycleGAN with shape and cycle consistency, three different loss types instead of the original two are combined during training:

Adversarial GAN losses \mathcal{I}_{GAN} . This loss term is the same as defined, e.g., in Eq. 5.

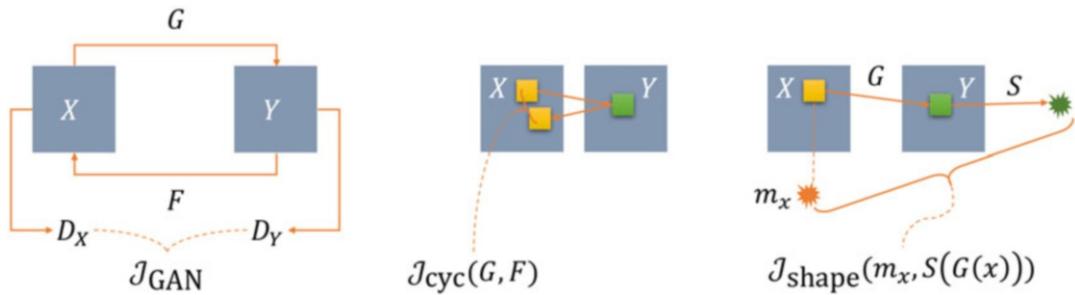


Fig. 20 Cycle GAN with shape consistency loss (rightmost part of figure). Note that the figure shows only one direction to ease readability

Cycle consistency losses \mathcal{J}_{cyc} .

This is the ℓ_1 loss presented by the original CycleGAN authors discussed above.

Shape consistency losses $\mathcal{J}_{\text{shape}}$.

The shape consistency loss is a new addition proposed by the authors. A cross-correlation loss takes into account two segmentations, the first being the gold standard segmentation m_x for an $x \in X$ and one segmentation produced by a segmenter network S that was trained on domain Y and receives the translated image $\hat{y} = G(x)$.

Figure 20 depicts the three loss components, of which the first two are known already from Fig. 19.

Note that the description as well as Fig. 20 only show one direction for cycle and shape consistency loss. Both are duplicated into the other direction and combined into the overall training objective, which then consists of six components.

In several other works, the CycleGAN approach was extended and combined with domain adaption methods for various segmentation tasks and also extended to volumetric data [41–43].

4.3 StyleGAN and Successor

One of the most powerful image synthesis GANs to date is the successor of StyleGAN, StyleGAN2 [44, 45]. The authors, at the time of writing researching at Nvidia, deviate from the usual GAN approach in which an image is generated from a randomly sampled vector from a latent space. Instead, they use a latent space that is created by a mapping function f which is in their architecture implemented as a multilayer perceptron which maps from a 512-dimensional space Z into a 512-dimensional space W . The second major change consisted of the so-called adaptive instance normalization layer, AdaIN, which implements a normalization to zero-mean and unit variance of each feature map, followed by a multiplicative factor and an additive bias term. This serves to

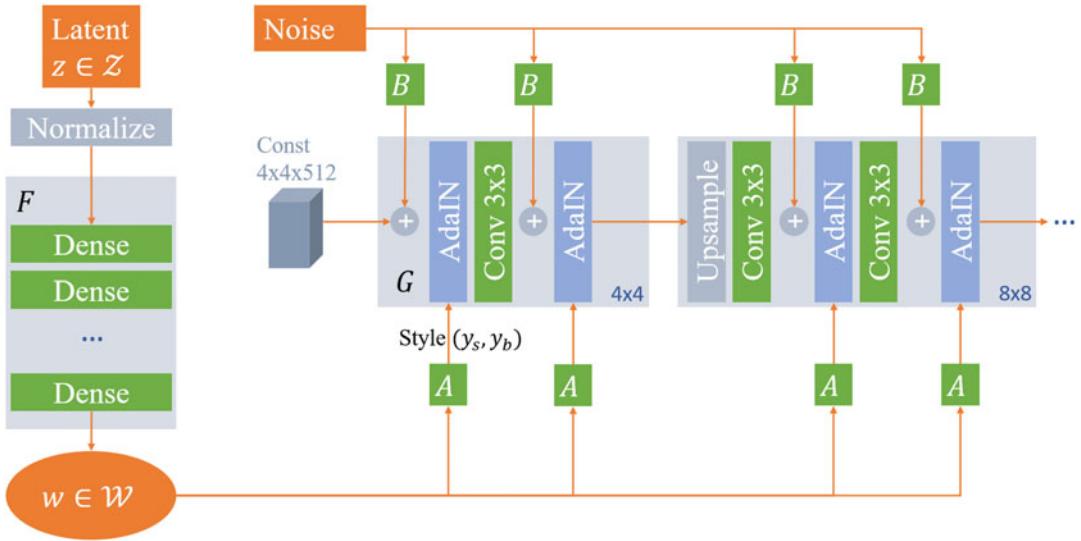


Fig. 21 StyleGAN architecture, after [44]. Learnable layers and transformations are shown in green, the AdaIN function in blue

reweight the importance of feature maps in one layer. To ensure the locality of the reweighting, the operation is followed by the non-linearity. The scaling and bias are two components of $\mathbf{y} = (y_s, y_b)$, which is the result of a learnable affine transformation A applied to a sample from W .

In their experiments, Karras et al. [44] recognized that after these changes, the GAN actually no longer depended on the input vector drawn from W itself, so the random latent vector was replaced by a static vector fed into the GAN. The y , which they call *styles*, remained to be results from a vector randomly sampled from the new embedding space W .

Lastly, noise is added in each layer, which serves to allow the GAN to produce more variation without learning to produce it from actual image content. The noise, like the latent vector, is fed through learnable transformations B , before it is added to the unnormalized feature maps. The overall architecture is sketched in Fig. 21.

In the basic setup, one sample is drawn from W and fed through per-layer learned A to gain per-layer different interpretations of the style, $\mathbf{y} = (y_s, y_b)$. This can be changed, however, and the authors show how using one random sample w_1 in some of the layer blocks and another sample w_2 in the remaining; the result will be a mixture of styles of both individual samples. This way, the coarse attributes of the generated image can stem from one sample and the fine detail from another. Applied to a face generator, for example, pose and shape of the face are determined in the coarse early layers of the network, while hair structure and skin texture are the fine



Fig. 22 Images created with StyleGAN; <https://thispersondoesnotexist.com>. Last accessed: 2022-01-14

details of the last layers. The architecture and results gained widespread attention through a website,¹⁰ which recently was followed up by further similar pages. Results are depicted in Fig. 22.

The crucial finding in StyleGAN was that the mapping function F transforming the latent space vector from Z to W serves to ensure a disentangled (flattened) latent space. Practically, this means that if interpolating points z_i between two points z_1 and z_2 drawn from Z and reconstructing images from these interpolated points z_i , semantic objects might appear (in a StyleGAN-generating faces, for example, a hat or glasses) that are neither part of the generated images from the first point z_1 nor the second point z_2 between which it has been interpolated. Conversely, if interpolating in W , this “semantic discontinuity” is no longer the case, as the authors show with experiments in which they measure the visual change of resulting images when traversing both latent spaces.

In their follow-up publications, the same authors improve the performance even further. They stick to the basic architecture but redesign the generative network pertaining to the AdaIN function. In addition, they add their metric from [44] that was meant to quantify the entanglement of the latent space as a regularizer. The discriminator network was also enhanced, and the mechanisms of StyleGAN that implement the progressive growing have been successively replaced by more performance-efficient setups. In their experiments, they show a growth of visual and measured quality and removal of several artifacts reported for StyleGAN [45].

4.4 Stabilized GAN for Few-Shot Learning

GAN training was very demanding both regarding GPU power, in particular for high-performance architectures like StyleGAN and StyleGAN2, and, as importantly, availability of data. StyleGAN2, for example, has typical training times of about 10 days on a Nvidia 8-GPU Tesla V100. The datasets comprised at least tens of thousands of images and easily orders of magnitude more. Particularly in the medical domain, such richness of data is typically hard to find.

¹⁰ <https://thispersondoesnotexist.com/>.

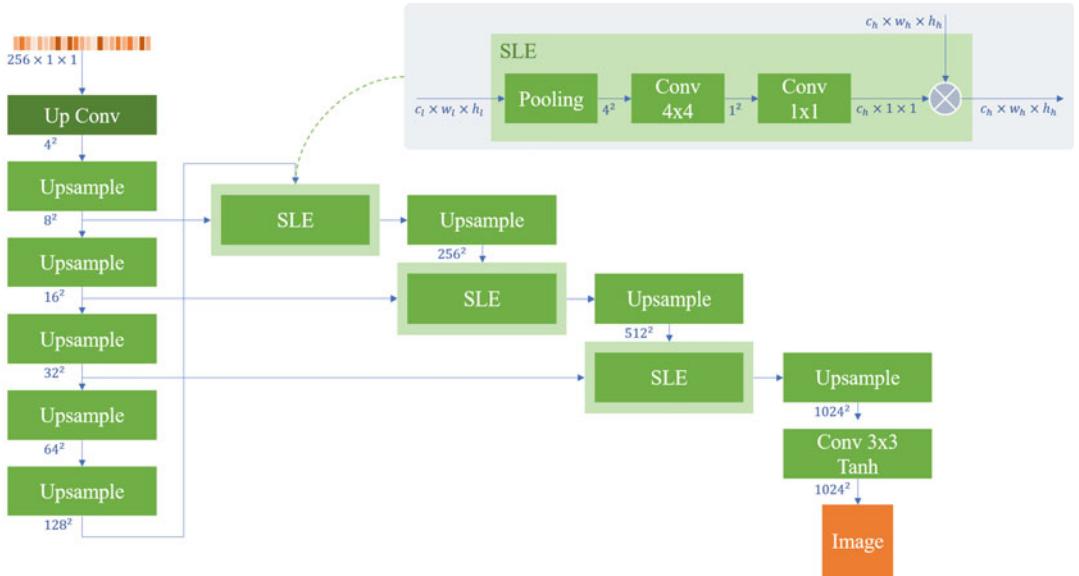


Fig. 23 The FastGAN generator network. Shortcut connections through feature map weighting layers (called skip-layer excitation, SLE) transport information from low-resolution feature maps into high-resolution feature maps. For details regarding the blocks, see text

The authors of [46] propose simple measures to stabilize the training of a specific GAN architecture, which they design from scratch using a replacement for residual blocks, arranged in an architecture with very few convolutional layers, and a loss that drives the discriminator to be less certain when it gets closer to convergence. In sum, this achieves very fast training and yields results competitive with prior GANs [46] and outperforming them in low-data situations.

The key ingredients to the architecture are shortcut connections in the generator model that rescale feature maps of higher resolution with learnable weights derived from low resolutions. The effect is to make fine details simultaneously more independent of direct predecessor feature maps and yet ensure consistency across scales.

A random seed vector of length 256 enters the first block (“Up Conv”), where it is upscaled to a $256 \times 4 \times 4$ tensor. In Fig. 23, the further key blocks of the architecture are “upsample” and “SLE” blocks.

Upsample blocks consist of a nearest-neighbor upsampling followed by a 3×3 convolution, batch normalization, and nonlinearity.

SLE blocks (seen in the top right inset in the architecture diagram) don’t touch the incoming high-resolution input (entering from top into the block) but comprise a pooling layer that in each SLE block is set up to yield a

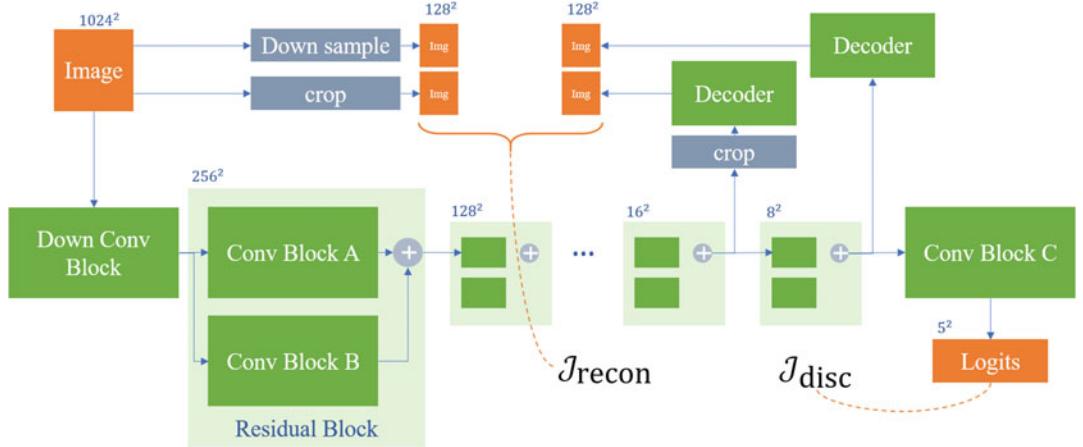


Fig. 24 The FastGAN self-supervision mechanism of the discriminator network. Self-supervision manifests through the loss term indicated by the curly bracket between reconstructions from feature maps and resampled/cropped versions of the original real image, J_{recon}

4 × 4 stack of feature maps, followed by a convolution to reduce to a 1 × 1 tensor, which is then in a 1 × 1 convolution brought to the same number of channels as the high-resolution input. This vector is then multiplied to the channels of the high-resolution input.

Secondly, the architecture introduces a self-supervision feature in the discriminator network. The discriminator network (*see* Fig. 24) is a simple CNN with strided convolutions in each layer, halving resolution in each feature map. In the latest (coarsest) feature maps, simple up-scaling convolutional networks are attached that generate small images, which are then compared in loss functions (J_{recon} in Fig. 24) to down-sampled versions of the real input image. This self-supervision of the discriminator is only performed for real images, not for generated ones.

The blocks in the figure spell out as follows:

- | | |
|--|--|
| Down Conv Block
Residual Blocks | consists of two convolutional layers with strided 4 × 4 convolutions, effectively reducing the resolution from 1024^2 to 256^2 .
have two sub-items, “Conv Block A” being a strided 4 × 4 convolution to half resolution, followed by a padded 3 × 3 convolution. “Conv Block B” consists of a strided 2 × 2 average pooling that quarters resolution, followed by a 1 × 1 convolution, so that both blocks result in identically shaped tensors, which are then added. |
|--|--|

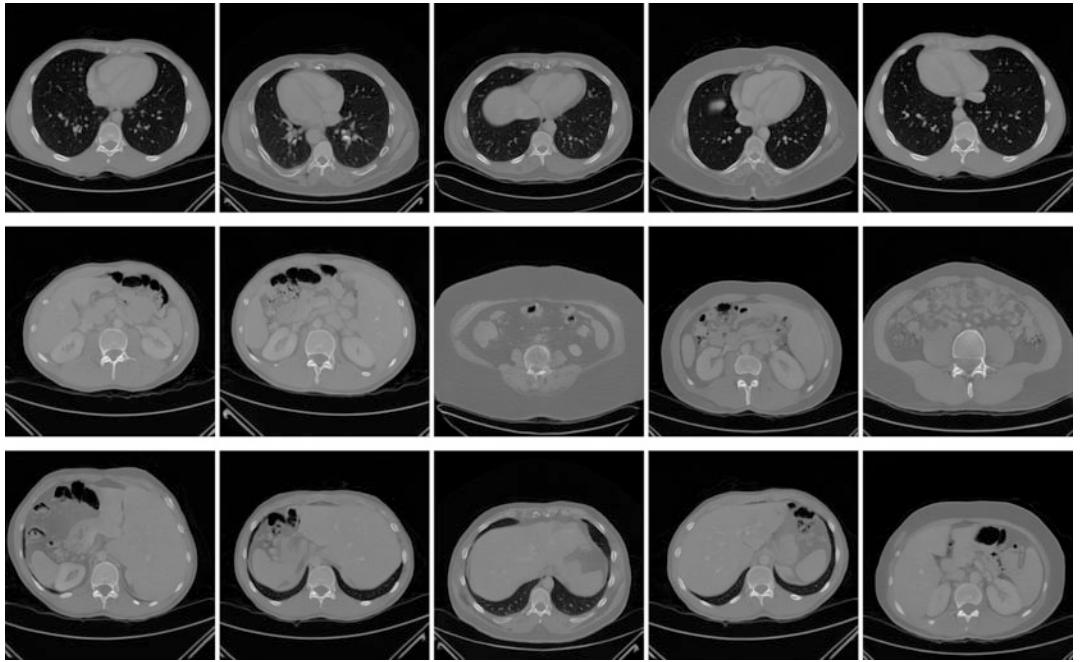


Fig. 25 FastGAN as implemented by the authors has been used to train a CT slice generative model. Images are not cherry-picked, but arranged by similar anatomical regions

Conv Block C

consists of a 1×1 convolution followed by a 4×4 convolution without strides or padding, so that the incoming 8^2 feature map is reduced to 5^2 .

Decoder

The decoder networks are four blocks of upsampling layers each followed by 3×3 convolutions.

The losses employed in the model are the discriminator loss consisting of the hinge version of the usual GAN loss, with the added regularizing reconstruction loss between original real samples and their reconstruction, and the generator loss plainly being $\mathcal{I}_G = \mathbb{E}_{z \sim Z}[D(G(z))]$.

The model is easy to train on modest hardware and little data, as evidenced by own experiments on a set of about 30 chest CTs (about 2500 image slices, converted to RGB). Figure 25 shows randomly picked generated example slices, roughly arranged by anatomical content. It is to be noted that organs appear mirrored in some images. On the other hand, no color artifacts are visible, so that the model has learned to produce only gray scale images. Training time for 50,000 iterations on a Nvidia TitanX GPU was approximately 10 hours.



Fig. 26 The VQGAN+CLIP combination creates images from text inputs, here: “A child drawing of a dark garden full of animals”

4.5 VQGAN

In a recent development, a team of researchers combined techniques for text interpretation with a dictionary of elementary image elements feeding into a generative network. The basic architecture component that is employed goes back to vector quantization variational autoencoders (VQ-VAE), where the latent space is no longer allowed to be continuous, but is quantized. This allows to use the latent space vectors in a look-up table: the visual elements.

Figure 26 was created using code available [online](#), which demonstrates how images of different visual styles can be created using the combination of text-based conditioning and a powerful generative network.

The basis for image generation is the VQGAN (“vector quantization generative adversarial network”) [47], which learns representations of input images that can later steer the generative process, in an adversarial framework. The conditioning is achieved with the CLIP (“Contrastive Image-Language Pretraining”) model that learns a discriminator that can judge plausible images for a text label or vice versa [48].

The architecture has been developed with an observation in mind that puts the benefits and drawbacks of convolutional and transformer architectures in relation to each other. While the locality bias of convolutional architectures is inappropriate if overall structural image relations should be considered, it is of great help in capturing textural details that can exist anywhere, like fur, hair, pavement, or grass, but where the exact representation of hair

positions or pavement stones is irrelevant. On the other hand, image transformers are known to learn convolutional operators implicitly, posing a severe computational burden without a visible impact on the results. Therefore, Esser et al. [47] suggest to combine convolutional operators for local detail representation and transformer-based components for image structure.

Since the VQGAN as a whole is no longer a pure CNN but for a crucial component uses a transformer architecture, this model will be brought up again briefly in Subheading 5.2.

The VQGAN architecture is derived from the VQ-VAE (vector quantization variational autoencoder) [49], adding a reconstruction loss through a discriminator, which turns it into a GAN. At the core of the architecture is the quantization of estimated codebook entries. Among the quantized entries in the codebook, the closest entry to the query vector coding, an image patch is determined. The found codebook entry is then referred to by its index in the codebook. This quantization operation is non-differentiable, so for end-to-end training, gradients are simply copied through it during backpropagation.

The transformer can then efficiently learn to predict codebook indices from those comprising the current version of the image, and the generative part of the architecture, the decoder, produces a new version of the image. Learning expressive codebook entries is enforced by a perceptual loss that punishes inaccurate local texture, etc. Through this, the authors can show that high compression levels can be achieved—a prerequisite to enable efficient, yet comprehensive, transformer training.

5 Other Generative Models

We have already seen how GANs were not the first approach to image generation but have prevailed for a time when they became computationally feasible and in consequence have been better understood and improved to accomplish tasks in image analysis and image generation with great success. In parallel with GANs, other fundamentally different generative modeling approaches have also been under continued development, most of which have precursors from the “before-GAN” era as well. To give a comprehensive outlook, we will sketch in this last section the state of the art of a selection of these approaches.¹¹

¹¹The research on the so-called flow-based models, e.g., normalizing flows, has been omitted in this chapter, though acknowledging their emerging relevance also in the context of image generation. Flow-based models are built from sequences of invertible transformations, so that they learn data distributions explicitly at the expense of sometimes higher computational costs due to their sequential architecture. When combined, e.g., with a powerful GAN, they allow innovative applications, for example, to steer the exploration of a GAN’s latent space to achieve fine-grained control over semantic attributes for conditional image generation. Interested readers are referred to the literature [11, 13, 50–52].

5.1 Diffusion and Score-Based Models

Diffusion models take a completely different approach to distribution estimation. GANs implicitly represent the target distribution by learning a surrogate distribution. Likelihood-based models like VAE approximate the target distribution explicitly, not requiring the surrogate. In diffusion models, however, the gradient of the log probability density function is estimated, instead of looking at the distribution itself (which would be the unfathomable integral of the gradient). This value is known as the Stein score function, leading to the notion that diffusion models are one variant of score-based models [53].

The simple idea behind this class of models is to revert a sequential noising process. Consider some image. Then, perform a large number of steps. In each step, add a small amount of noise from a known distribution, e.g., the normal distribution. Do this until the result is indistinguishable from random noise.

The denoising process is then formulated as a latent variable model, where $T - 1$ latents successively progress from a noise image $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ to the reconstruction that we call $\mathbf{x}_0 \sim q(\mathbf{x}_0)$. The reconstructed image, \mathbf{x}_0 , is therefore obtained by a *reverse process* $q_\theta(\mathbf{x}_{0:T})$. Note that each step in this chain can be evaluated in closed form [54]. Several model implementations of this approach exist, one being the deep diffusion probabilistic model (DDPM). Here, a deep neural network learns to perform one denoising step given the so-far achieved image and a $t \in \{1, \dots, T\}$. Iterative application of the model to the result of the last iteration will eventually yield a generated image from noise input.

Autoregressive diffusion models (ARDMs) [55] follow yet another thought model, roughly reminiscent of PixelRNNs we have briefly mentioned above (see Subheading 3.2). Both share the approach to condition the prediction of the next pixel or pixels on the already predicted ones. Other than in the PixelRNN, however, the specific ARDM proposed by the authors does not rely on a predetermined schedule of pixel updates, so that these models can be categorized as latent variable models.

As of late, the general topic of score-based methods, among which diffusion models are one variant, received more attention in the research community, fueled by a growing body of publications that report image synthesis results that outperform GANs [53, 56, 57]. Score function-based and diffusion models superficially share the similar concept of sequentially adding/removing noise but achieve their objective with very different means: where score function-based approaches are trained by score-matching and their sampling process uses Langevin dynamics [58], diffusion models are trained using the evidence lower bound (ELBO) and sample with a decoder, which is commonly a neural network. Figure 27 visualizes an example for a score function.

Score function-based (sometimes also score-matching) generative models have been developed to astounding quality levels, and

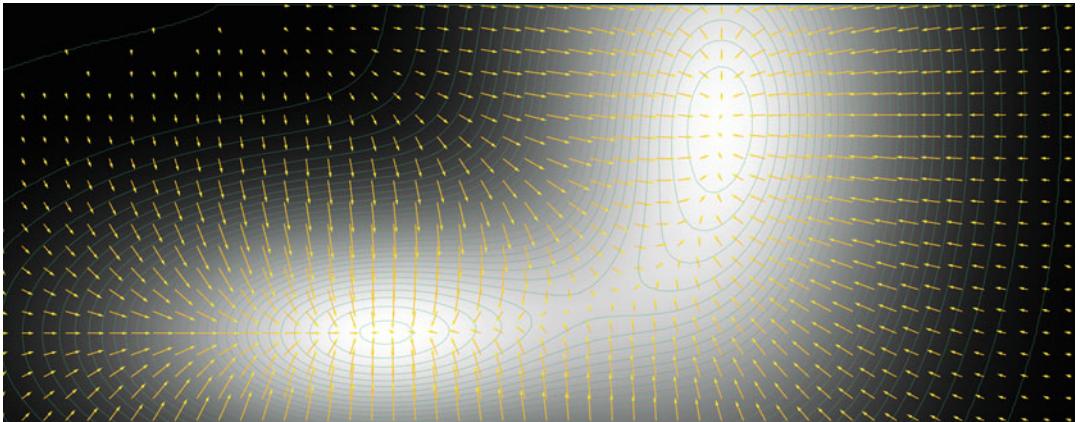


Fig. 27 The Stein score function can be conceived of as the gradient of the log probability density function, here indicated by two Gaussians. The arrows represent the score function

the recent works of Yang Song and others provide accessible blog posts,¹² and a comprehensive treatment of the subject in several publications [53, 58, 59].

In the work of Ho et al. [54], the stepwise reverse (denoising) process is the basis of the denoising diffusion probabilistic models (DDPM). The authors emphasize that a proper selection of the noise schedule is crucial to fast, yet high-quality, results. They point out that their work is a combination of diffusion probabilistic models with score-matching models, in this combination also generalizing and including the ideas of autoregressive denoising models. In an extension of Ho et al.’s [54] work by Nichol and Dhariwal [57], an importance sampling scheme was introduced that lets the denoising process steer the most easy to predict next image elements. Equipped with this new addition, the authors can show that, in comparison to GANs, a wider region of the target distribution is covered by the generative model.

5.2 Transformer-Based Generative Models

The basics of how attention mechanisms and transformer architectures work will be covered in the subsequent chapter on this promising technology (Chapter 6). Attention-based models, predominantly transformers, have been used successfully for some time in sequential data processing and are now considered the superior alternative to recurrent networks like long-short-term memory (LSTM) networks. Transformers have, however, only recently made their way into the image analysis and now also the image generation world. In this section, we will only highlight some developments in the area of generative tasks.

¹² <https://yang-song.github.io/blog/>.

Google Brain/Google AI’s 2018 publication on so-called image transformers [60], among other tasks, shows successful conditional image generation for low-resolution input images to achieve super-resolution output images, and for image inpainting, where missing or removed parts of input images are replaced by content produced by the image transformer.

OpenAI have later shown that even unmodified language transformers can succeed to model image data, by dealing in sheer compute power for hand modeling of domain knowledge, which was the basis for the great success of previous unsupervised image generation models. They have trained Image GPT (or iGPT for short), a multibillion parameter language transformer model, and it excels in several image generation tasks, though only for fairly small image sizes [61].

In the recent past, StyleSwin has been proposed by Microsoft Research Asia [62], enabling high-resolution image generation. However, the approach uses a block-wise attention window, thereby potentially introducing spatial incoherencies at block edges, which they have to correct for.

“Taming transformers” [47], another recent publication already mentioned above, uses what the authors call a learned template code book of image components, which is combined with a vector quantization GAN (VQGAN). The VQGAN is structurally modeled after the VQ-VAE but adds a discriminator network. A transformer model in this architecture composes these code book elements and is interrogated by the GAN variational latent space, conditioned on a textual input, a label image, or other possible inputs. The GAN reconstructs the image from the so-quantized latent space using a combination of a perceptual loss assessing the overall image structure and a patch-based high-resolution reconstruction loss. By using a sliding attention window approach, the authors prevent patch border artifacts known from StyleSwin. Conditioning on textual input makes use of parts of the CLIP [48] idea (“Contrastive Language-Image Pretraining”), where a language model was train in conjunction with an image encoder to learn embeddings of text-image pairs, sufficient to solve many image understanding tasks with competitive precision, without specific domain adaption.

It is evidenced by the lineup of institutions that training image transformer models successfully is nothing that can be achieved with modest hardware or on even a medium-scale image database. In particular for the medical area, where data is comparatively scarce even under best assumptions, the power of such models will only be available in the near future if domain transfer learning can be successfully achieved. This, however, is a known strength of transformer architectures.

Acknowledgements

I thank my colleague at the Fraunhofer Institute for Digital Medicine MEVIS, Till Nicke, for his thorough review of the chapter and many valuable suggestions for improvements. I owe many thanks more to other colleagues for their insights both in targeted discussions and most importantly in everyday work life.

References

- [1] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Proceedings of the 27th international conference on neural information processing systems - volume, NIPS'14 . MIT Press, Cambridge, pp 2672–2680
- [2] Casella G, Berger RL (2021) Statistical inference. Cengage Learning, Boston
- [3] Grinstead C, Snell LJ (2006) Introduction to probability. Swarthmore College, Swarthmore
- [4] Severini TA (2005) Elements of distribution theory, vol 17. Cambridge University Press, Cambridge
- [5] Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press, Cambridge
- [6] Murphy KP (2022) Probabilistic machine learning: an introduction. MIT Press, Cambridge. <http://doi.org/probml.ai>
- [7] Do CB, Batzoglou S (2008) What is the expectation maximization algorithm? Nat Biotechnol 26:8, 26:897–899. <https://doi.org/10.1038/nbt1406>. <https://www.nature.com/articles/nbt1406>
- [8] Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. J Roy Statist Soc Ser B (Methodolog) 39:1–22. <https://doi.org/10.1111/J.2517-6161.1977.TB01600.X>.<https://onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1977.tb01600.x>. <https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1977.tb01600.x>
- [9] van den Oord A, Kalchbrenner N, Kavukcuoglu K (2016) Pixel recurrent neural networks. ArXiv abs/1601.06759
- [10] Magnusson K (2020) Understanding maximum likelihood: an interactive visualization. <https://rpsychologist.com/likelihood/>
- [11] Rezende DJ, Mohamed S (2015) Variational inference with normalizing flows. In: ICML
- [12] van den Oord A, Kalchbrenner N, Espeholt L, Kavukcuoglu K, Vinyals O, Graves A (2016) Conditional image generation with PixelCNN decoders. In: NIPS
- [13] Dinh L, Sohl-Dickstein J, Bengio S (2017) Density estimation using Real NVP. ArXiv abs/1605.08803
- [14] Salakhutdinov R, Hinton G (2009) Deep Boltzmann machines. In: van Dyk D, Welling M (eds) Proceedings of the twelfth international conference on artificial intelligence and statistics, PMLR, hilton clearwater beach resort, clearwater beach, Florida USA, Proceedings of Machine Learning Research, vol 5, pp 448–455. <https://proceedings.mlr.press/v5/salakhutdinov09a.html>
- [15] Weng L (2018) From autoencoder to Beta-VAE. lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html
- [16] Kingma DP, Welling M (2014) Auto-encoding variational bayes. ArXiv [1312.6114](https://arxiv.org/abs/1312.6114)
- [17] Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA (2018) Generative adversarial networks: an overview. IEEE Signal Process Mag 35(1): 53–65. <https://doi.org/10.1109/MSP.2017.2765202>
- [18] Arjovsky M, Bottou L (2017) Towards principled methods for training generative adversarial networks. ArXiv abs/1701.04862
- [19] Theis L, van den Oord A, Bethge M (2016) A note on the evaluation of generative models. CoRR abs/1511.01844
- [20] Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with

- deep convolutional generative adversarial networks. ArXiV <http://arxiv.org/abs/1511.06434>
- [21] Islam J, Zhang Y (2020) GAN-based synthetic brain PET image generation. *Brain Inform* 7:1–12. <https://doi.org/10.1186/S40708-020-00104-2/FIGURES/9>. <https://braininformatics.springeropen.com/articles/10.1186/s40708-020-00104-2>
- [22] Arjovsky M, Chintala S, Bottou L (2017) Wasserstein GAN. ArXiV <http://arxiv.org/abs/1701.07875v3>. 1701.07875
- [23] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A (2017) Improved training of Wasserstein GANs. ArXiV <http://arxiv.org/abs/1704.00028v3>. nIPS camera-ready, 1704.00028
- [24] Villani C (2009) Optimal transport, old and new. Springer, Berlin. <https://doi.org/10.1007/978-3-540-71050-9>. <https://www.cedricvillani.org/wp-content/uploads/2012/08/preprint-1.pdf>
- [25] Basso G (2015) A Hitchhiker’s guide to Wasserstein distances. <https://homeweb.unifr.ch/BassoG/pub/A%20Hitchhikers%20guide%20to%20Wasserstein.pdf>
- [26] Weng L (2019) From GAN to WGAN. ArXiV 1904.08994
- [27] Baumgartner CF, Koch LM, Tezcan KC, Ang JX, Konukoglu E (2018) Visual feature attribution using Wasserstein GANs. In: The IEEE conference on computer vision and pattern recognition (CVPR)
- [28] Dzanic T, Shah K, Witherden FD (2020) Fourier spectrum discrepancies in deep network generated images. In: 34th conference on neural information processing systems (NeurIPS)
- [29] Joslin M, Hao S (2020) Attributing and detecting fake images generated by known GANs. In: Proceedings - 2020 IEEE symposium on security and privacy workshops, SPW 2020. Institute of Electrical and Electronics Engineers, Piscataway, pp 8–14. <https://doi.org/10.1109/SPW50608.2020.00019>
- [30] Le BM, Woo SS (2021) Exploring the asynchronous of the frequency spectra of GAN-generated facial images. ArXiV <https://arxiv.org/abs/2112.08050v1>. 2112.08050
- [31] Goebel M, Nataraj L, Nanjundaswamy T, Mohammed TM, Chandrasekaran S, Manjunath BS, Maya (2021) Detection, attribution and localization of GAN generated images. *Electron Imag*. <https://doi.org/10.2352/ISSN.2470-1173.2021.4.MWSF-276>
- [32] Isola P, Zhu JY, Zhou T, Efros AA (2016) Image-to-image translation with conditional adversarial networks. ArXiV <http://arxiv.org/abs/1611.07004>
- [33] Senaras C, Sahiner B, Tozbikian G, Lozanski G, Gurcan MN (2018) Creating synthetic digital slides using conditional generative adversarial networks: application to Ki67 staining. In: Medical imaging 2018: digital pathology, society of photo-optical instrumentation engineers (SPIE) conference series, vol 10581, p 1058103. <https://doi.org/10.1117/12.2294999>
- [34] Zhao G, Meyerand ME, Birn RM (2021) Bayesian conditional GAN for MRI brain image synthesis. ArXiV 2005.11875
- [35] Bakas S, Reyes M, ..., Menze B (2019) Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. ArXiV 1811.02629
- [36] Cirillo MD, Abramian D, Eklund A (2020) Vox2Vox: 3D-GAN for brain tumour segmentation. ArXiV 2003.13653
- [37] Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE international conference on computer vision (ICCV), IEEE, pp 2242–2251. <http://ieeexplore.ieee.org/document/8237506/papers3://publication/doi/10.1109/ICCV.2017.244>
- [38] Almahairi A, Rajeswar S, Sordoni A, Bachman P, Courville A (2018) Augmented CycleGAN: Learning many-to-many mappings from unpaired data. ArXiV <https://arxiv.org/pdf/1802.10151.pdf>. 1802.10151
- [39] Chu C, Zhmoginov A, Sandler M (2017) CycleGAN, a master of steganography. ArXiV <http://arxiv.org/abs/1712.02950>
- [40] Zhang Z, Yang L, Zheng Y (2018) Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, IEEE, pp 9242–9251. <https://doi.org/10.1109/CVPR.2018.00963>. <https://ieeexplore.ieee.org/document/8579061/>

- [41] Hoffman J, Tzeng E, Park T, Zhu JY, Isola P, Saenko K, Efros AA, Darrell T (2017) CyCADA: Cycle-consistent adversarial domain adaptation. ArXiV [1711.03213](#)
- [42] Huo Y, Xu Z, Bao S, Assad A, Abramson RG, Landman BA (2018) Adversarial synthesis learning enables segmentation without target modality ground truth. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), pp 1217–1220. <https://doi.org/10.1109/ISBI.2018.8363790>
- [43] Yang D, Xiong T, Xu D, Zhou SK (2020) Segmentation using adversarial image-to-image networks. In: Handbook of medical image computing and computer assisted intervention, pp 165–182. <https://doi.org/10.1016/B978-0-12-816176-0.00012-0>
- [44] Karras T, Laine S, Aila T (2018) A style-based generator architecture for generative adversarial networks. IEEE Trans Pattern Anal Mach Intell 43:4217–4228. <https://doi.org/10.1109/TPAMI.2020.2970919>. <https://arxiv.org/abs/1812.04948v3>
- [45] Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 8107–8116. <https://doi.org/10.1109/CVPR42600.2020.00813>. <https://arxiv.org/abs/1912.04958v2>
- [46] Liu B, Zhu Y, Song K, Elgammal A (2021) Towards faster and stabilized GAN training for high-fidelity few-shot image synthesis. In: International conference on learning representations. <https://openreview.net/forum?id=1Fqg133qRaI>
- [47] Esser P, Rombach R, Ommer B (2021) Taming transformers for high-resolution image synthesis. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 12868–12878. <https://doi.org/10.1109/CVPR46437.2021.01268>
- [48] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning transferable visual models from natural language supervision. ArXiV [2103.00020](#)
- [49] van den Oord A, Vinyals O, Kavukcuoglu K (2017) Neural discrete representation learning. CorRR abs/1711.00937. <http://arxiv.org/abs/1711.00937>
- [50] Weng L (2018) Flow-based deep generative models. lilianwenggithubio/lil-log. <http://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html>
- [51] Kingma DP, Dhariwal P (2018) Glow: generative flow with invertible 1x1 convolutions. ArXiv <https://doi.org/10.48550/ARXIV.1807.03039>. <https://arxiv.org/abs/1807.03039>
- [52] Abdal R, Zhu P, Mitra NJ, Wonka P (2021) StyleFlow: attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. ACM Trans Graph 40(3):1–21. <https://doi.org/10.1145/3447648>. <https://doi.org/10.1145%2F3447648>
- [53] Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B (2021) Score-based generative modeling through stochastic differential equations. In: International conference on learning representations. <https://openreview.net/forum?id=PxtIG12RRHS>
- [54] Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. ArXiV [2006.11239](#)
- [55] Hoogeboom E, Gritsenko AA, Bastings J, Poole B, van den Berg R, Salimans T (2021) Autoregressive diffusion models. ArXiV [2110.02037](#)
- [56] Dhariwal P, Nichol A (2021) Diffusion models beat GANs on image synthesis. ArXiV <http://arxiv.org/abs/2105.05233>
- [57] Nichol A, Dhariwal P (2021) Improved denoising diffusion probabilistic models. ArXiV <http://arxiv.org/abs/2102.09672>
- [58] Song Y, Ermon S (2019) Generative modeling by estimating gradients of the data distribution. In: Advances in neural information processing systems, pp 11895–11907
- [59] Song Y, Garg S, Shi J, Ermon S (2019) Sliced score matching: a scalable approach to density and score estimation. In: Proceedings of the thirty-fifth conference on uncertainty in artificial intelligence, UAI 2019, Tel Aviv, Israel, July 22–25, 2019, p 204. <http://auai.org/uai2019/proceedings/papers/204.pdf>

- [60] Parmar N, Vaswani A, Uszkoreit J, Łukasz Kaiser, Shazeer N, Ku A, Tran D (2018) Image transformer. ArXiv [1802.05751](https://arxiv.org/abs/1802.05751)
- [61] Chen M, Radford A, Child R, Wu J, Jun H, Luan D, Sutskever I (2020) Generative pre-training from pixels. In: Daumé III H, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, proceedings of machine learning research, vol 119, pp 1691–1703. <https://proceedings.mlr.press/v119/chen20s.html>
- [62] Zhang B, Gu S, Zhang B, Bao J, Chen D, Wen F, Wang Y, Guo B (2021) StyleSwin: transformer-based GAN for high-resolution image generation. ArXiv [2112.10762](https://arxiv.org/abs/2112.10762)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 6

Transformers and Visual Transformers

Robin Courant, Maika Edberg, Nicolas Dufour, and Vicky Kalogeiton

Abstract

Transformers were initially introduced for natural language processing (NLP) tasks, but fast they were adopted by most deep learning fields, including computer vision. They measure the relationships between pairs of input tokens (words in the case of text strings, parts of images for visual transformers), termed attention. The cost is exponential with the number of tokens. For image classification, the most common transformer architecture uses only the transformer encoder in order to transform the various input tokens. However, there are also numerous other applications in which the decoder part of the traditional transformer architecture is also used. Here, we first introduce the attention mechanism (Subheading 1) and then the basic transformer block including the vision transformer (Subheading 2). Next, we discuss some improvements of visual transformers to account for small datasets or less computation (Subheading 3). Finally, we introduce visual transformers applied to tasks other than image classification, such as detection, segmentation, generation, and training without labels (Subheading 4) and other domains, such as video or multimodality using text or audio data (Subheading 5).

Key words Attention, Transformers, Visual transformers, Multimodal attention

1 Attention

Attention is a technique in Computer Science that imitates the way in which the brain can focus on the relevant parts of the input. In this section, we introduce attention: its history (Subheading 1.1), its definition (Subheading 1.2), its types and variations (Subheadings 1.3 and 1.4), and its properties (Subheading 1.5).

To understand what attention is and why it is so useful, consider the following film review:

While others claim the story is boring, I found it fascinating.

Is this film review positive or negative? The first part of the sentence is unrelated to the critic's opinion, while the second part suggests a positive sentiment with the word 'fascinating'. To a human, the answer is obvious; however, this type of analysis is not necessarily obvious to a computer.

Typically, sequential data require *context* to be understood. In natural language, a word has a meaning because of its position in the sentence, with respect to the other words: its *context*. In our example, while “boring” alone suggests that the review is negative, its contextual relationship with other words allows the reader to reach the appropriate conclusion. In computer vision, in a task like object detection, the nature of a pixel alone cannot be identified: we need to account for its neighborhood, its *context*. So, how can we formalize the concept of *context* in sequential data?

1.1 The History of Attention

This notion of *context* is the motivation behind the introduction of the attention mechanism in 2015 [1]. Before this, language translation was mostly relying on encoder-decoder architectures: recurrent neural networks (RNNs) [2] and in particular long-short-term memory (LSTMs) networks were used to model the relationship among words [3]. Specifically, each word of an input sentence is processed by the encoder sequentially. At each step, the past and present information are summarized and encoded into a fixed-length vector. In the end, the encoder has processed every word and outputs a final fixed-length vector, which summarizes all input information. This final vector is then decoded and finally translates the input information into the target language.

However, the main issue of such structure is that all the information is compressed into one fixed-length vector. Given that the sizes of sentences vary and as the sentences get longer, a fixed-length vector is a real bottleneck: it gets increasingly difficult not to lose any information in the encoding process due to the vanishing gradient problem [1].

As a solution to this issue, Bahdanau et al. [1] proposed the attention module in 2015. The attention module allows the model to consider the parts of the sentence that are relevant to predicting the next word. Moreover, this facilitates the understanding of relationships among words that are further apart.

1.2 Definition of Attention

Given two lists of tokens, $X \in \mathbb{R}^{N \times d_x}$ and $\Upsilon \in \mathbb{R}^{N \times d_y}$, attention encodes information from Υ into X , where N is the length of inputs X and Υ and d_x and d_y are their respective dimensions. For this, we first define three linear mappings, query mapping $W^Q \in \mathbb{R}^{d_x \times d_q}$, key mapping $W^K \in \mathbb{R}^{d_y \times d_k}$, and value mapping $W^V \in \mathbb{R}^{d_y \times d_v}$, where d_q , d_k , and d_v are the embedding dimensions in which the query, key, and value are going to be computed, respectively.

Then, we define the query Q , key K , and value V [4] as:

$$\begin{aligned} Q &= X W^Q \\ K &= \Upsilon W^K \\ V &= \Upsilon W^V \end{aligned}$$

Next, the *attention matrix* is defined as:

$$A(\mathbf{Q}, \mathbf{K}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right). \quad (1)$$

This is illustrated in the left part of Fig. 1. The nominator $\mathbf{Q}\mathbf{K}^T \in \mathbb{R}^{N \times N}$ represents how each part of the input in \mathbf{X} attends to each part of the input in \mathbf{Y} .¹ This dot product is then put through the softmax function to normalize its values and get positive values that add to 1. However, for large values of d_k , this may result in the softmax to have incredibly small gradients, so it is scaled down by $\sqrt{d_k}$.

The resulting $N \times N$ matrix encodes the relationship between \mathbf{X} with respect to \mathbf{Y} : it measures how important a token in \mathbf{X} is with respect to another one in \mathbf{Y} .

Finally, the *attention output* is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = A(\mathbf{Q}, \mathbf{K})\mathbf{V}. \quad (2)$$

Figure 1 displays this. The attention output encodes the information of each token by taking into account the contextual information. Therefore, through the learnable parameters—queries, keys, and values—the attention layers learn a token embedding that takes into account their relationship.

Contextual Relationships How does Eq. 2 encode contextual relationships? To answer this question, let us reconsider analyzing the sentiment of film reviews. To encode contextual relationships into the word embedding, we first want a matrix representation of the relationship between all words. To do so, given a sentence of length N , we take each word vector and feed it to two different linear layers, calling one output “query” and the other output “key”. We pack the queries into the matrix \mathbf{Q} and the keys into the matrix \mathbf{K} , by taking their product ($\mathbf{Q}\mathbf{K}^T$). The result is a $N \times N$ matrix that explains how important the i -th word (row-wise) is to understand the j -th word (column-wise). This matrix is then scaled and normalized by the division and softmax. Next, we feed the word vectors into another linear layer, calling its output “value”. We multiply these two matrices together. The results of their product are attention vectors that encode the meaning of each word, by including their contextual meaning as well. Given that each of these queries, keys, and values is learnable parameter, as the attention layer is trained, the model learns how relationships among words are encoded in the data.

¹ Note that in the literature, there are two main attention functions: additive attention [1] and dot-product attention (Eq. 1). In practice, the dot product is more efficient since it is implemented using highly optimized matrix multiplication, compared to the feed-forward network of the additive attention; hence, the dot product is the dominant one.

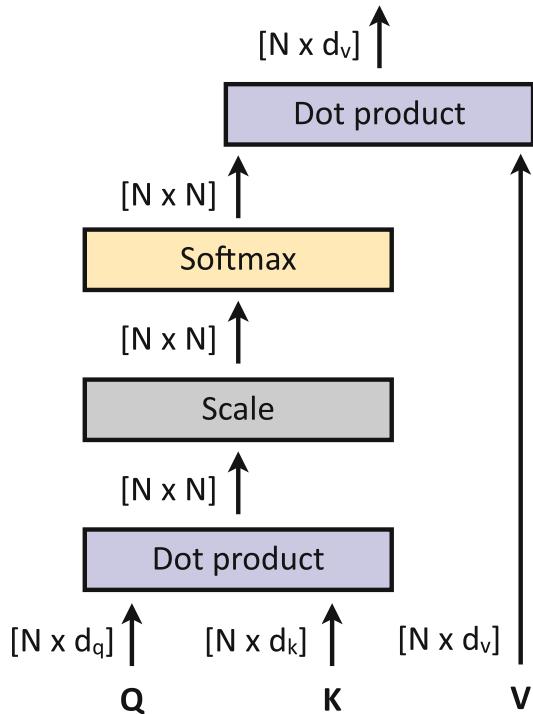


Fig. 1 Attention block. Next to each element, we denote its dimensionality. Figure inspired from [4]

1.3 Types of Attention

There exist two dominant types of attention mechanisms: *self-attention* and *cross attention* [4]. In *self-attention*, the queries, keys, and values come from the same input, i.e., $X = Y$; in *cross attention*, the queries come from a different input than the key and value vectors, i.e., $X \neq Y$. These are described below in Subheadings 1.3.1 and 1.3.2, respectively.

1.3.1 Self-Attention

In self-attention, the tokens of X attend to themselves ($X = Y$). Therefore, it is modeled as follows:

$$\text{SA}(X) = \text{Attention}(XW^Q, XW^K, XW^V). \quad (3)$$

Self-attention formalizes the concept of context. It learns the patterns underlying how parts of the input correspond to each other. By gathering information from the same set, given a sequence of tokens, a token can attend to its neighboring tokens to compute its output.

1.3.2 Cross Attention

Most real-world data are multimodal—for instance, videos contain frames, audios, and subtitles, images come with captions, etc. Therefore, models that can deal with such types of multimodal information have become essential.

Cross attention is an attention mechanism designed to handle multimodal inputs. Unlike self-attention, it extracts queries from one input source and key-value pairs from another one ($X \neq Y$). It answers the following question: “Which parts of input X and input Y correspond to each other?” Cross attention (CA) is defined as:

$$\text{CA}(X, Y) = \text{Attention}(XW^Q, YW^K, YW^V). \quad (4)$$

1.4 Variation of Attention

Attention is typically employed in two ways: (1) multi-head self-attention (MSA, Subheading 1.4) and (2) masked multi-head attention (MMA, Subheading 1.4).

Attention Head We call attention head the mechanism presented in Subheading 1.2, i.e., query-key-value projection, followed by scaled dot product attention (Eqs. 1 and 2).

When employing an attention-based model, relying only on a single attention head can inhibit learning. Therefore, the multi-head attention block is introduced [4].

Multi-head Self-Attention (MSA) MSA is shown in Fig. 2 and is defined as:

$$\begin{aligned} \text{MSA}(X) &= \text{Concat}(\text{head}_1(X), \dots, \text{head}_h(X))W^O, \\ \text{head}_i(X) &= \text{SA}(X), \forall i \in \{1, h\}, \end{aligned} \quad (5)$$

where Concat is the concatenation of h attention heads and $W^O \in \mathbb{R}^{hd_v \times d}$ is projection matrix. This means that the initial embedding dimension d_x is decomposed into $h \times d_v$ and the computation per head is carried out independently. The independent attention heads are usually concatenated and multiplied by a linear layer to match the desired output dimension. The output dimension is often the same as the input embedding dimension d . This allows an easier stacking of multiple blocks.

Multi-head Cross Attention (MCA) Similar to MSA, MCA is defined as:

$$\begin{aligned} \text{MCA}(X, Y) &= \text{Concat}(\text{head}_1(X, Y), \dots, \text{head}_h(X, Y))W^O, \\ \text{head}_i(X, Y) &= \text{CA}(X, Y), \forall i \in \{1, h\}. \end{aligned} \quad (6)$$

Masked Multi-head Self-Attention (MMSA) The MMSA layer [4] is another variation of attention. It has the same structure as the multi-head self-attention block (Subheading 1.4), but all the later vectors in the target output are masked. When dealing with sequential data, this can help make training parallel.

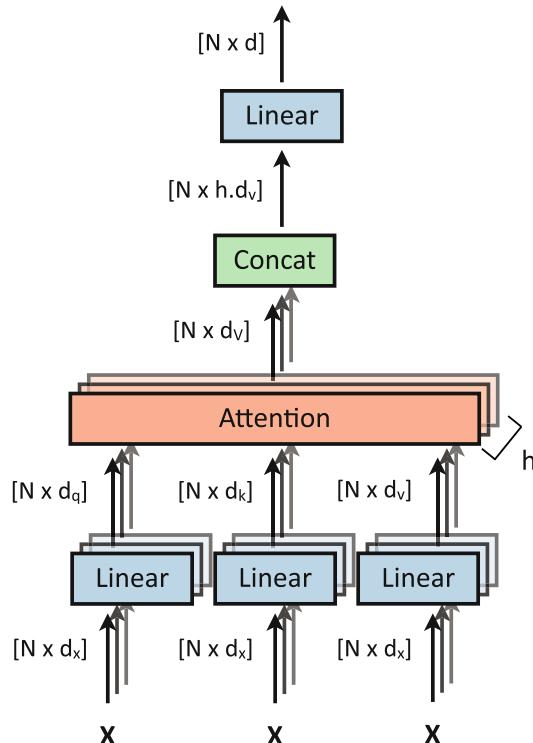


Fig. 2 Multi-head self-attention block (MSA). First, the input \mathbf{X} is projected to queries, keys, and values and then passed through h attention blocks. The h resulting attention outputs are then concatenated together and finally projected to a d -dimensional output vector. Next to each element, we denote its dimensionality. Figure inspired from [4]

1.5 Properties of Attention

While attention encodes contextual relationships, it is *permutation equivalent*, as the mechanism does not account for the order of the input data. As shown in Eq. 2, the attention computations are all matrix multiplication and normalizations. Therefore, a permuted input results in a permuted output. In practice, however, this may not be an accurate representation of the information. For instance, consider the sentences “the monkey ate the banana” and “the banana ate the monkey.” They have distinct meanings because of the order of the words. If the order of the input is important, various mechanisms, such as the positional encoding, discussed in Subheading 2.1.2, are used to capture this subtlety.

2 Visual Transformers

The transformer architecture was introduced in [4] and is the first architecture that relies purely on attention to draw connections between the inputs and outputs. Since its debut, it revolutionized deep learning, making breakthroughs in numerous fields, including

natural language processing, computer vision, chemistry, and biology, thus making its way to becoming the *default* architecture for learning representations. Recently, the standard transformer [4] has been adapted for vision tasks [5]. And again, visual transformer has become one of the central architectures in computer vision.

In this section, we first introduce the basic architecture of transformers (Subheading 2.1) and then present its advantages (Subheading 2.2). Finally, we describe the vision transformer (Subheading 2.3).

2.1 Basic Transformers

As shown in Fig. 3, the transformer architecture [4] is an encoder-decoder model. First, it embeds input tokens $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ into a latent space, resulting in latent vectors $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)$, which are fed to the decoder to output $\mathbf{Y} = (y_1, \dots, y_M)$. The encoder is a stack of L layers, with each one consisting of two sub-blocks: multi-head self-attention (MSA) layers and a multilayer perceptron (MLP). The decoder is also a stack of L layers, with each one consisting of three sub-blocks: masked multi-head self-attention (MMSA), multi-head cross attention (MCA), and MLP.

Overview Below, we describe the various parts of the transformer architecture, following Fig. 3. First, the input tokens are converted into the embedding tokens (Subheading 2.1.1). Then, the positional encoding adds a positional token to each embedding token to denote the order of tokens (Subheading 2.1.2). Then, the transformer encoder follows (Subheading 2.1.3). This consists of a stack of L multi-head attention, normalization, and MLP layers and encodes the input to a set of semantically meaningful features. After, the decoder follows (Subheading 2.1.4). This consists of a stack of L masked multi-head attention, multi-head attention, and MLP layers followed by normalizations and decodes the input features with respect to the output embedding tokens. Finally, the output is projected to linear and softmax layers.

2.1.1 Embedding

The first step of transformers consists in converting input tokens² into embedding tokens, i.e., vectors with meaningful features. To do so, following standard practice [6], each input is projected into an embedding space to obtain embedding tokens \mathbf{Z}^e . The embedding space is structured in a way that the distance between a pair of vectors is relative to the semantic similarity of their associated words. For the initial NLP case, this means that we get a vector of each word, such that the vectors that are closer together have similar meanings.

² Note the initial transformer architecture was proposed for natural language processing (NLP), and therefore the inputs were words.

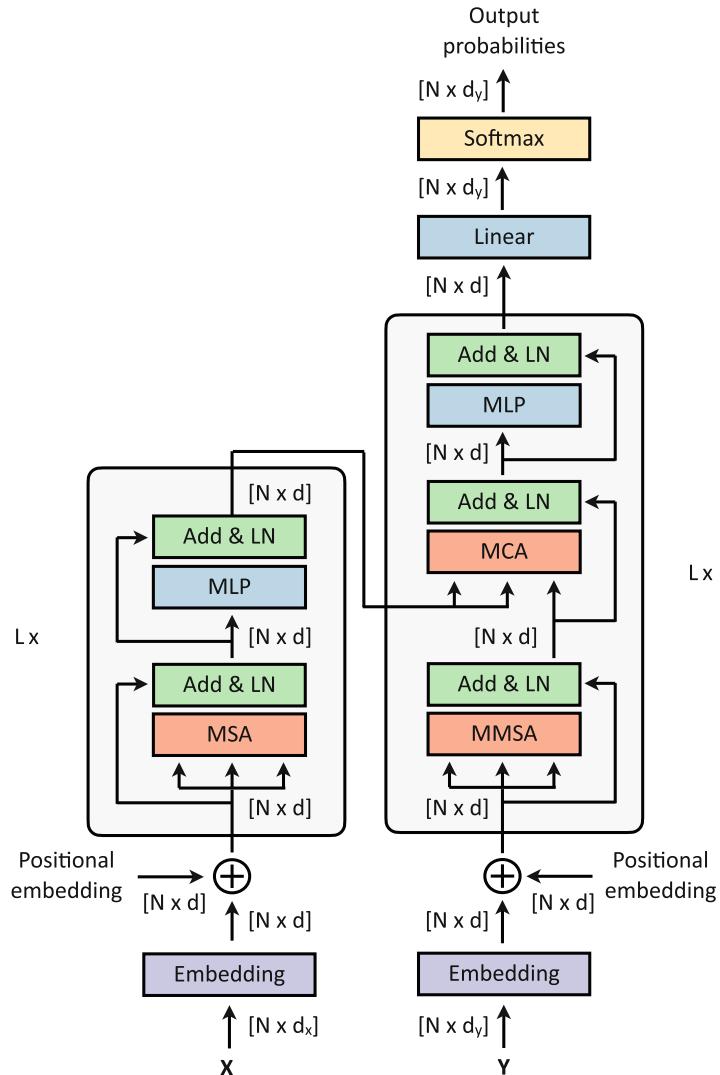


Fig. 3 The transformer architecture. It consists of an encoder (left) and a decoder (right) block, each one consisting from a series of attention blocks (multi-head and masked multi-head attention) and MLP layers. Next to each element, we denote its dimensionality. Figure inspired from [4]

2.1.2 Positional Encoding

As discussed in Subheading 1.5, the attention mechanism is positional agnostic, which means that it does not store the information on the position of each input. However, in most cases, the order of input tokens is relevant and should be taken into account, such as the order of words in a sentence matter as they may change its meaning. Therefore, [4] introduced the *Positional Encoding* $PE \in \mathbb{R}^{N \times d_x}$, which adds a positional token to each embedding token $Z^e \in \mathbb{R}^{N \times d_x}$.

Sinusoidal Positional Encoding

The sinusoidal positional encoding [4] is the main positional encoding method, which encodes the position of each token with sinusoidal waves of multiple frequency. For an embedding token $\mathbf{Z}^e \in \mathbb{R}^{N \times d_x}$, its positional encoding $\mathbf{PE} \in \mathbb{R}^{N \times d_x}$ is defined as:

$$\begin{aligned}\mathbf{PE}(i, 2j) &= \sin\left(\frac{i}{10000^{2j/d}}\right) \\ \mathbf{PE}(i, 2j + 1) &= \cos\left(\frac{i}{10000^{2j/d}}\right), \forall i, j \in [|1, n|] \times [|1, d|].\end{aligned}\tag{7}$$

Learnable Positional Encoding

An orthogonal approach is to let the model learn the positional encoding. In this case, $\mathbf{PE} \in \mathbb{R}^{N \times d_x}$ becomes a learnable parameter. This, however, increases the memory requirements, without necessarily bringing improvements over the sinusoidal encoding.

Positional Embedding

After its computation, either the positional encoding \mathbf{PE} is added to the embedding tokens or they are concatenated as follows:

$$\begin{aligned}\mathbf{Z}^{\text{pe}} &= \mathbf{Z}^e + \mathbf{PE}, \text{ or} \\ \mathbf{Z}^{\text{pe}} &= \text{Concat}(\mathbf{Z}^e, \mathbf{PE}),\end{aligned}\tag{8}$$

where Concat denotes vector concatenation. Note that the concatenation has the advantage of not altering the information contained in \mathbf{Z}^e , since the positional information is only added to the unused dimension. Nevertheless, it augments the input dimension, leading to higher memory requirements. Instead, the addition does preserve the same input dimension while altering the content of the embedding tokens. When the input dimension is high, this content altering is trivial, as most of the content is preserved. Therefore, in practice, for high dimension, summing positional encodings is preferred, whereas for low dimensions concatenating them prevails.

2.1.3 Encoder Block

The encoder block takes as input the embedding and positional tokens and outputs features of the input, to be decoded by the decoder block. It consists of a stack of L multi-head self-attention (MSA) layers and a multilayer perceptron (MLP). Specifically, the embedding and positional tokens, $\mathbf{Z}_x^{\text{pe}} \in \mathbb{R}^{N \times d}$, go through a multi-head self-attention block. Then, a residual connection with layer normalization is deployed. In the transformer, this operation is performed after each sub-layer. Next, we feed its output to an MLP and a normalization layer. This operation is performed L times, and each time the output of each encoder block (of size $N \times d$) is the input of the subsequent block. In the L -th time, the output of the normalization is the input of the cross-attention block in the decoder (Subheading 2.1.4).

2.1.4 Decoder Block

The decoder has two inputs: first, an input that constitutes the queries $\mathbf{Q} \in \mathbb{R}^{N \times d}$ of the encoder, and, second, the output of the encoder that constitutes the key-value $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times d}$ pair. Similar to Subheadings 2.1.1 and 2.1.2, the first step constitutes encoding the output token to output embedding token and output positional token. These tokens are fed into the main part of the decoder, which consists of a stack of L masked multi-head self-attention (MMSA) layers, multi-head cross-attention (MCA) layers, and multilayer perceptron (MLP) followed by normalizations. Specifically, the embedding and positional tokens, $\mathbf{Z}_y^{\text{pe}} \in \mathbb{R}^{N \times d}$, go through a MMSA block. Then, a residual connection with layer normalization follows. Next, an MCA layer (followed by normalization) maps the queries to the encoded key values before forwarding the output to an MLP. Finally, we project the output of the L decoder blocks (of dimension $N \times d_y$) through a linear layer and get output probability through a softmax layer.

2.2 Advantages of Transformers

Since their introduction, the transformers have had a significant impact on deep learning approaches.

In natural language processing (NLP), before transformers, most architectures used to rely on recurrent modules, such as RNNs [2] and in particular LSTMs [3]. However, recurrent models process the input sequentially, meaning that, to compute the current state, they require the output of the previous state. This makes them tremendously inefficient, as they are impossible to parallelize. On the contrary, in transformers, each input is processed independent of the others, and the multi-head attention can perform multiple attention computations at once. This makes transformers highly efficient, as they are highly parallelizable.

This results in not only exceptional scalability, both in the complexity of the model and the size of datasets, but also relatively fast training. Notably, the recent switch transformers [7] was pre-trained on 34 billion tokens from the C4 dataset [8], scaling the model to over 1 trillion parameters.

This scalability [7] is the principal reason for the power of the transformer. While it was originally introduced for translation, it refrains from introducing many inductive biases, i.e., the set of assumptions that the user makes about the structure of the model input. In doing so, the transformer relies on data to learn how they are structured. Compared to its counterparts with more biases, the transformer requires much more data to produce comparable results [5]. However, if a sufficient amount of data is available, the lack of inductive bias becomes a strength. By learning the structure of the data from the data, the transformer is able to learn better without human assumptions hindering [9].

In most tasks involving transformers, the model is first pre-trained on a large dataset and then fine-tuned for the task at hand on a smaller dataset. The pretraining phase is essential for

transformers to learn the global structure of the specific input modality. For fine-tuning, typically fewer data suffice as the model is already rich. For instance, in natural language processing, BERT [10], a state-of-the-art language model, is pretrained on a Wikipedia-based dataset [11], with over 6 million articles and Book Corpus [12] with over 10,000 books. Then, this model can be fine-tuned on much more specific tasks. In computer vision, the vision transformer (ViT) is pretrained on the JFT-300M dataset, containing over 1 billion labels for 300 million images [5]. Hence, with a sufficient amount of data, transformers achieve results that were never possible before in various areas of machine learning.

2.3 Vision Transformer

Transformers offer an alternative to CNNs that have long held a stranglehold on computer vision. Before 2020, most attempts to use transformers for vision tasks were still highly reliant on CNNs, either by using self-attention jointly with convolutions [13, 14] or by keeping the general structure of CNNs while using self-attention [15, 16].

The reason for this is rooted in the two main weaknesses of the transformers. First, the complexity of the attention operation is high. As attention is a quadratic operation, the number of parameters skyrockets quickly when dealing with visual data, i.e., images—and even more so with videos. For instance, in the case of ImageNet [17], inputting a single image with $256 \times 256 = 65,536$ pixels in an attention layer would be too heavy computationally. Second, transformers suffer from lack of inductive biases. Since CNNs were specifically created for vision tasks, their architecture includes spatial inductive biases, like translation equivariance and locality. Therefore, the transformers have to be pretrained on a significantly large dataset to achieve similar performances.

The vision transformer (ViT) [5] is the first systematic approach that uses directly transformers for vision tasks by addressing both aforementioned issues. It rids the concept of convolutions altogether, using purely a transformer-based architecture. In doing so, it achieves the state of the art on image recognition on various datasets, including ImageNet [17] and CIFAR-100 [18].

Figure 4 illustrates the ViT architecture. The input image is first split into 16×16 patches, flattened, and mapped to the expected dimension through a learnable linear projection. Since the image size is reduced to 16×16 , the complexity of the attention mechanism is no longer a bottleneck. Then, ViT encodes the positional information and attaches a learnable embedding to the front of the sequence, similarly to BERT’s classification token [10]. The output of this token represents the entirety of the input—it encodes the information from each part of the input. Then, this sequence is fed into an encoder block, with the same structure as in the standard transformers [4]. The output of the classification token is then fed into an MLP that outputs class probabilities.

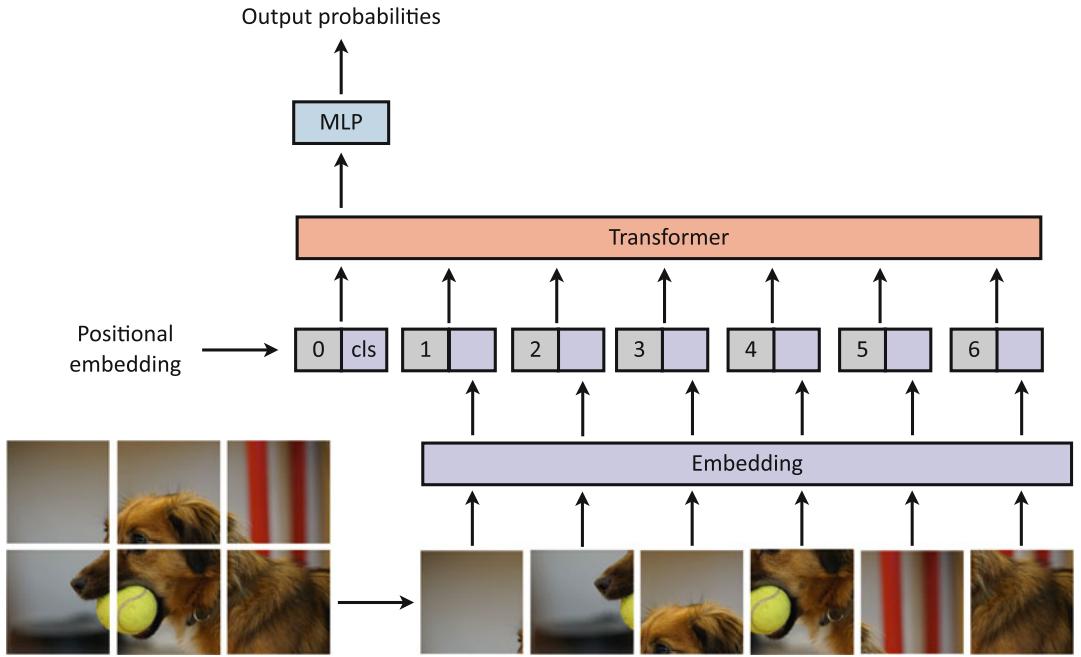


Fig. 4 The vision transformer architecture (ViT). First, the input image is split into patches (bottom), which are linearly projected (embedding), and then concatenated with positional embedding tokens. The resulting tokens are fed into a transformer, and finally the resulting classification token is passed through an MLP to compute output probabilities. Figure inspired from [5]

Due to the lack of inductive biases, when ViT is trained only on mid-sized datasets such as ImageNet, it scores some percentage points lower than the state of the art. Therefore, the proposed model is first pretrained on the JFT-300M dataset [19] and then fine-tuned on smaller datasets, thereby increasing its accuracy by 13%.

For a complete overview of visual transformers and follow-up works, we invite the readers to study [9, 20].

3 Improvements over the Vision Transformer

In this section, we present transformer-based methods that improve over the original vision transformer (Subheading 2.3) in two main ways. First, we introduce approaches that are trained on smaller datasets, unlike ViT [5] that requires pretraining on 300 million labeled images (Subheading 3.1). Second, we present extensions over ViT that are more computational-efficient than ViT, given that training a ViT is directly correlated to the image resolution and the number of patches (Subheading 3.2).

3.1 Data Efficiency

As discussed in Subheading 2.3, the vision transformer (ViT) [5] is pretrained on a massive proprietary dataset (JFT-300M) which contains 300 million labeled images. This need arises with transformers because we remove the inductive biases from the architecture compared to convolutional-based networks. Indeed, convolutions contain some translation equivariance. ViT does not benefit from this property and thus has to learn such biases, requiring more data. JFT-300M is an enormous dataset, and to make ViT work in practice, better data-efficiency is needed. Indeed, collecting that amount of data is costly and can be infeasible for most tasks.

Data-Efficient Image Transformers (DeiT) [21] The first work to achieve an improved data efficiency is DeiT [21]. The main idea of DeiT is to distil the inductive biases from a CNN into a transformer (Fig. 5). DeiT adds another token that works similarly to the class token. When training, ground truth labels are used to train the network according to the class token output with a cross-entropy (CE) loss. However, for the distillation network, the output labels are compared to the labels provided from a teacher network with a

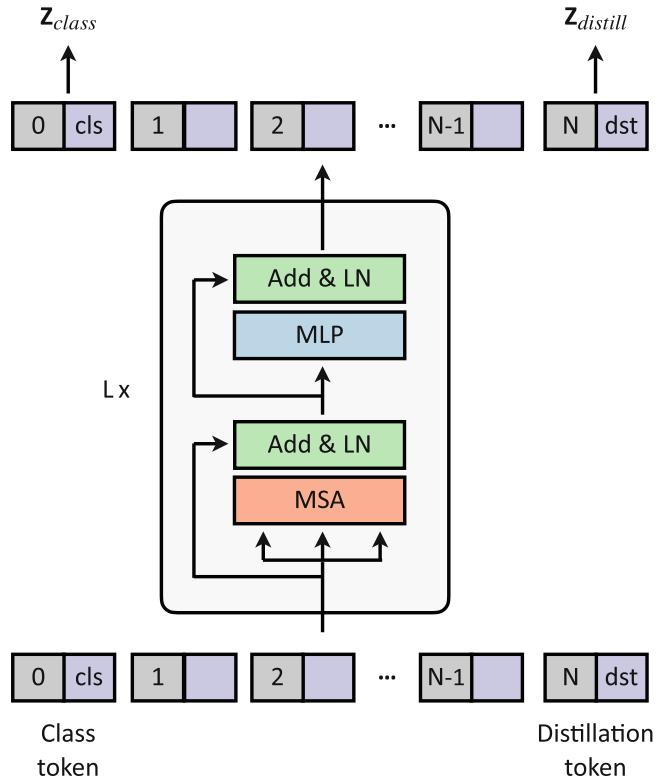


Fig. 5 The DeiT architecture. The architecture features an extra token, the distillation token. This token is used similarly to the class token. Figure inspired from [21]

cross-entropy loss. The final loss for a N -categorical classification task is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{global}}^{\text{hardDistill}} &= \frac{1}{2} (\mathcal{L}_{\text{CE}}(\Psi(\mathbf{Z}_{\text{class}}), \mathbf{y}) + \mathcal{L}_{\text{CE}}(\Psi(\mathbf{Z}_{\text{distill}}), \mathbf{y}_T)), \\ \mathcal{L}_{\text{CE}}(\hat{\mathbf{y}}, \mathbf{y}) &= -\frac{1}{N} \sum_{i=1}^N [y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)]\end{aligned}\quad (9)$$

with Ψ the softmax function, $\mathbf{Z}_{\text{class}}$ the class token output, $\mathbf{Z}_{\text{distill}}$ the class token output, \mathbf{y} the ground truth label, and \mathbf{y}_T the teacher label prediction.

The teacher network is a Convolutional Neural Network (CNN). The main idea is that the distillation head will provide the inductive biases needed to improve the data efficiency of the architecture. By doing this, DeiT achieves remarkable performance on the ImageNet dataset, by training “only” on ImageNet-1K [17], which contains 1.3 million images.

Convit [22] The main disadvantage of DeiT [21] is that it requires a pretrained CNN, which is not ideal, and it would be more convenient to not have this requirement. The CNN has a hard inductive bias constraint that can be a major limitation. Indeed, if enough data is available, learning the biases from the data can result in better representations.

Convit [22] overpasses this issue by including the inductive bias of CNNs into a transformer in a soft way. Specifically, if the inductive bias is limiting the training, the transformer can discard it. The main idea is to include the inductive bias into the ViT initialization. Therefore, before beginning training, the ViT is equivalent to a CNN. Then, the network can progressively learn the needed biases and diverge from the CNN initialization.

Compact Convolutional Transformer [23], DeiT [21], and Convit [22] successfully achieve data efficiency at the ImageNet scale. However, ImageNet is a big dataset with 1.3 million images, whereas most datasets are significantly smaller.

To reach higher data efficiency, the compact convolutional transformer [23] uses a CNN operation to extract the patches and then uses these patches in a transformer network (Fig. 6). The compact convolutional transformer comes with some modifications that lead to major improvements. First, by having a more complex encoding of patches, the system relies on the convolutional inductive biases at the lower scales and then uses a transformer network to remove the locality constraint of the CNN. Second, the authors show that discarding the “class” token results in higher efficiency. Specifically, instead of the class token, the compact convolutional transformer pools together all the patches token and classifies on top of this pooled token. These two modifications enable using

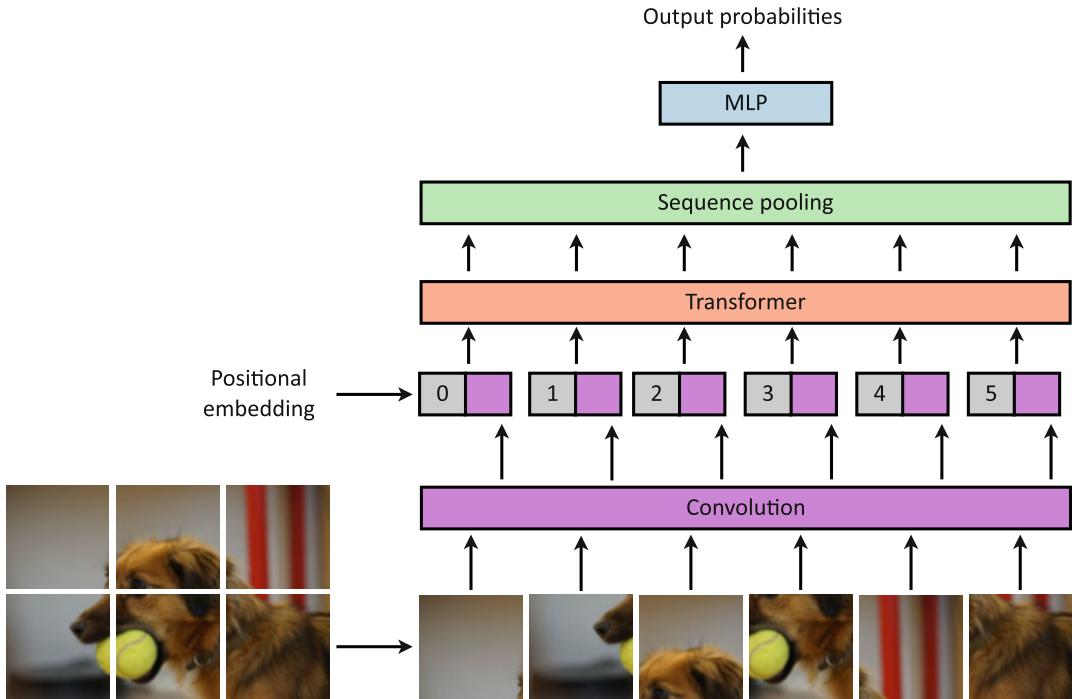


Fig. 6 Compact convolutional transformers. This architecture features a convolutional-based patch extraction to leverage a smaller transformer network, leading to higher data efficiency. Figure inspired from [23]

smaller transformers while improving both the data efficiency and the computational efficiency. Therefore, these improvements allow the compact convolutional transformer to be successfully trained on smaller datasets, such as CIFAR or MNIST.

3.2 Computational Efficiency

The vision transformer architecture (Subheading 2.3) suffers from a $\mathcal{O}(n^2)$ complexity with respect to the number of tokens. When considering small resolution images or big patch size, this is not a limitation; for instance, for an image of 224×224 resolution with 16×16 patches, this amounts to 196 tokens. However, when needing to process larger images (for instance, 3D images in medical imaging) or when considering smaller patches, using and training such models becomes *prohibitive*. For instance, in tasks such as segmentation or image generation, it is needed to have more granular representations than 16×16 patches; hence, it is crucial to solve this issue to enable more applications of vision transformer.

Swin Transformer [24] One idea to make transformers more computation-efficient is the Swin transformer [24]. Instead of attending every patch in the image, the Swin transformer proposes to add a locality constraint. Specifically, the patches can only attend other patches that are limited to a vicinity window K . This restores the local inductive bias of CNNs. To allow communication across

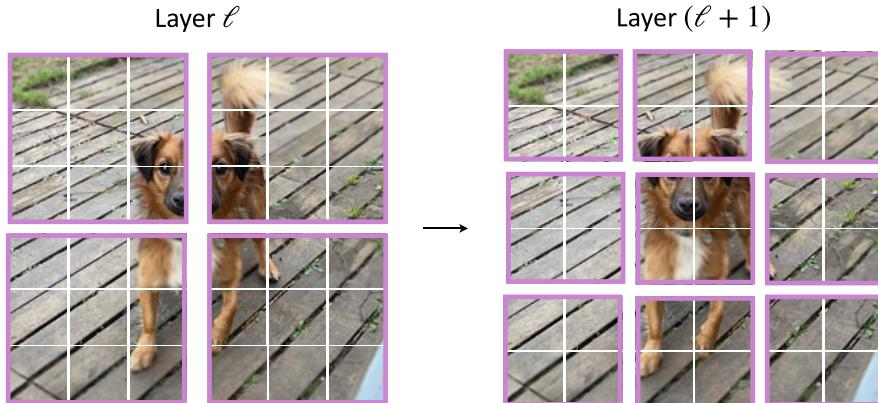


Fig. 7 Shifting operation in the Swin transformer [24]. Between each attention operation, the attention window is shifted so that each patch can communicate with different patches than before. This allows the network to gain more global knowledge with the network’s depth. Figure inspired from [24]

patches throughout the network, the Swin transformer shifts the attention windows from one operation to another (Fig. 7). Therefore, the Swin transformer is quadratic with regard to the size of the window K but linear with respect to the number of tokens n with complexity $\mathcal{O}(nK^2)$. In practice, however, K is small, and this solves the quadratic complexity problem of attention.

Perceiver [25, 26] Another idea for more computation-efficient visual transformers is to make a more drastic change to the architecture. If instead of using self-attention the model uses cross attention, the problem of the quadratic complexity with regard to the number of tokens can be solved. Indeed, computing the cross attention between two sets of length m and n , respectively, has complexity $\mathcal{O}(mn)$. This idea is introduced in the perceiver [25, 26]. The key idea is to have a smaller set of latent variables that will be used as queries and that will retrieve information in the image token set (Fig. 8). Since this solves the quadratic complexity issue, it also removes the need of using patches; hence, in the case of transformers, each pixel is mapped to a single token.

4 Vision Transformers for Tasks Other than Classification

Subheadings 1–3 introduce visual transformers for one main application: classification. Nevertheless, transformers can be used for numerous other tasks than classification.

In this section, we present some fundamental vision tasks where transformers have had a major impact: object detection in images (Subheading 4.1), image segmentation (Subheading 4.2), training

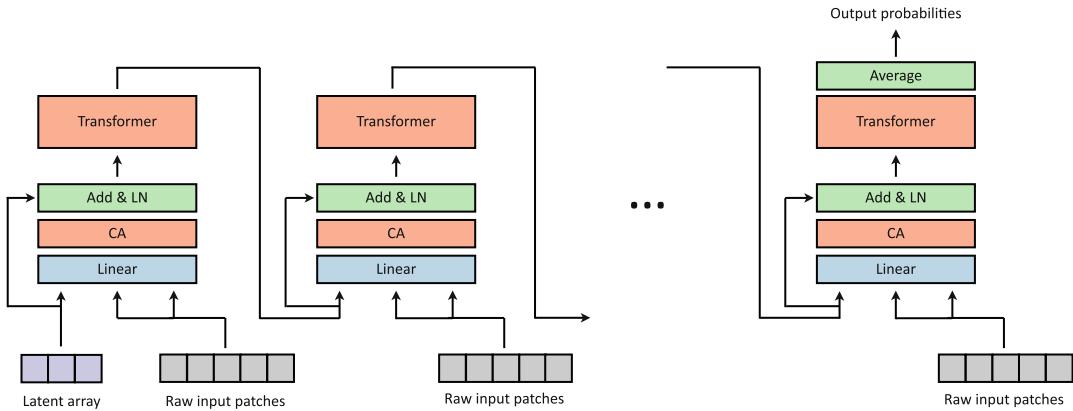


Fig. 8 The perceiver architecture [25, 26]. A set of latent tokens retrieve information from the image through cross attention. Self-attention is performed between the tokens to refine the learned representation. These operations are linear with respect to the number of image tokens. Figure inspired from [25, 26]

visual transformers without labels (Subheading 4.3), and image generation using generative adversarial networks (GANs) (Subheading 4.4).

4.1 Object Detection with Transformers

Detection is one of the early tasks that have seen improvements thanks to transformers. Detection is a combined recognition and localization problem; this means that a successful detection system should both recognize whether an object is present in an image and localize it spatially in the image. Carion et al. [14] is the first approach that uses transformers for detection.

DEtection TRansformer (DETR) [14] DETR first extracts visual representations with a convolutional network (Fig. 9).³ Then, the encodings are processed by a transformer network. Finally, the processed tokens are provided to a transformer decoder. The decoder uses cross attention between a set of learned tokens and the image tokens encoded by the encoder and outputs a set of tokens. Each output token is then passed through a feed-forward network that predicts if an object is present in an image or not; if the object is indeed present, the network also predicts the class and spatial location of the object, i.e., coordinates within the image.

4.2 Image Segmentation with Transformers

The goal of image segmentation is to assign to each pixel of an image the label of the object it belongs to. The segmenter [27] is a purely ViT approach addressing image segmentation. The idea is to first use ViT to encode the image. Then, the segmenter learns a token per

³ Note that, in DETR, the transformer is not directly used to extract the visual representation. Instead, it focuses on refining the visual representation to extract the object information.

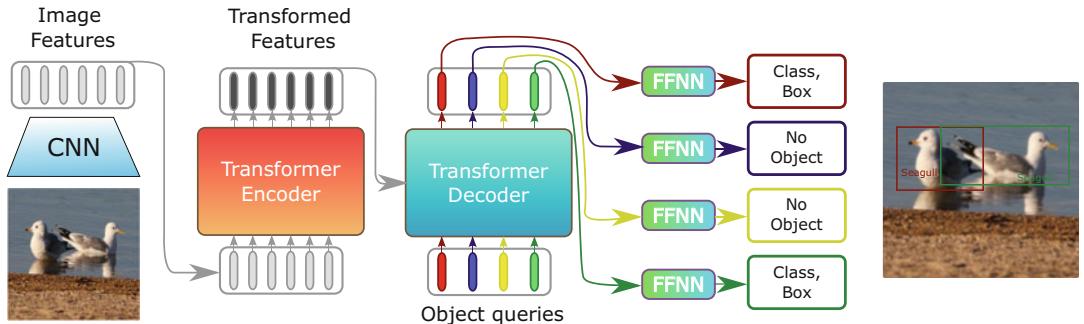


Fig. 9 The DETR architecture. It refines a CNN visual representation to extract object localization and classes. Figure inspired from [14]

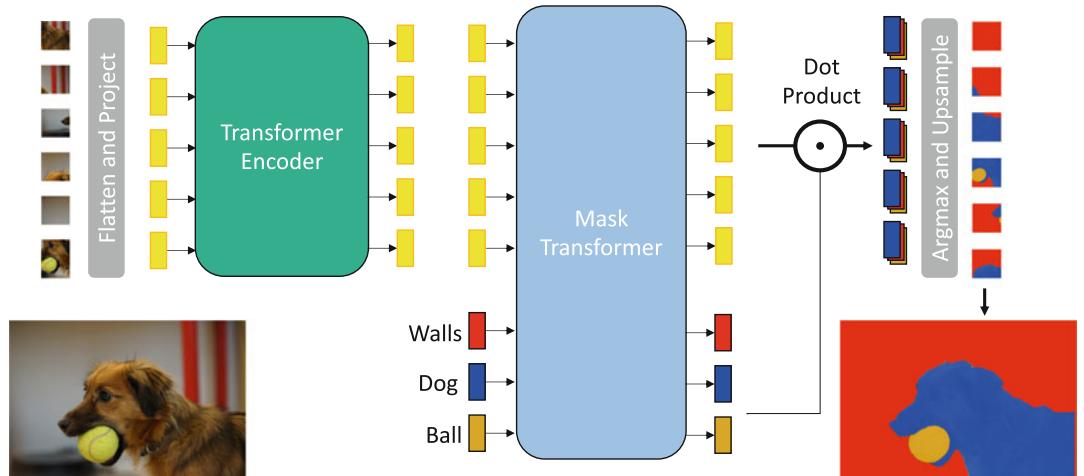


Fig. 10 The segmenter architecture. It is a purely ViT-based approach to perform semantic segmentation. Figure inspired from [27]

semantic label. The encoded patch tokens and the semantic tokens are then fed to a second transformer. Finally, by computing the scalar product between the semantic tokens and the image tokens, the network assigns a label to each patch. Figure 10 displays this.

4.3 Training Transformers Without Labels

Visual transformers have initially been trained for classification tasks. However, this task requires having access to massive amounts of labeled data, which can be hard to obtain (as discussed in Subheading 3.1). Subheadings 3.1 and 3.2 present ways to train ViT more efficiently. However, it would also be interesting to be able to train this type of networks with “cheaper” data. Therefore, the goal of this part is to introduce unsupervised learning with transformers, i.e., training transformers without any labels.

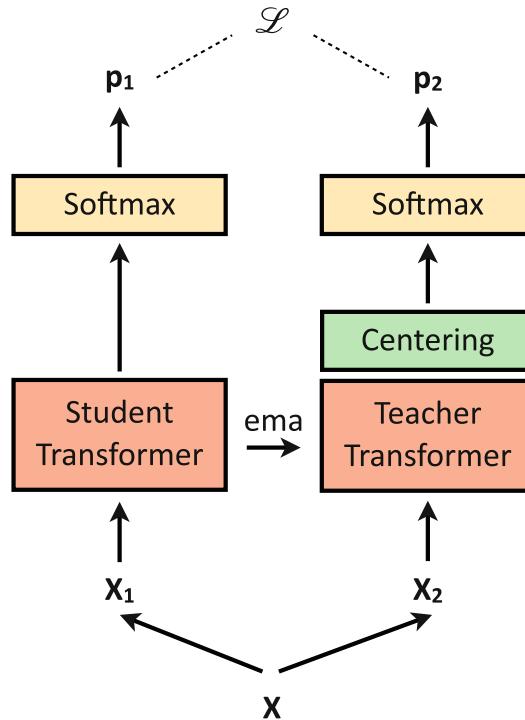


Fig. 11 The DINO training procedure. It consists in matching the outputs between two networks (p_1 and p_2) having two different augmentations (X_1 and X_2) of the same image as input (X). The parameters of the teacher model are updated with an exponential moving average (ema) of the student parameters. Figure inspired from [28]

Self-DIstillation with NO labels (DINO) [28] DINO is one of the first works that trains a ViT with self-supervised learning (Fig. 11). The main idea is to have two ViT models following the teacher-student paradigm: the first model is updated through gradient descent, and the second is an exponential moving average of the first one. Then, the whole two-stream DINO network is trained using two augmentations of the same image, which are each passed to one of the two networks. The goal of the training is to match the output between the two networks, i.e., no matter the augmentation in the input data, both networks should produce the same result. The main finding of DINO is that the ViT is capable of learning a semantic understanding of the image, as the attention matrices display some semantic information. Figure 12 visualizes the attention matrix of the various ViT heads trained with DINO.

Masked Autoencoders (MAE) [29] Another way to train a ViT without supervision is by using an autoencoder architecture. Masked autoencoders (MAE) [29] perform a random masking of the input token and give the task to reconstruct the original image to a decoder. The encoder learns a representation that performs

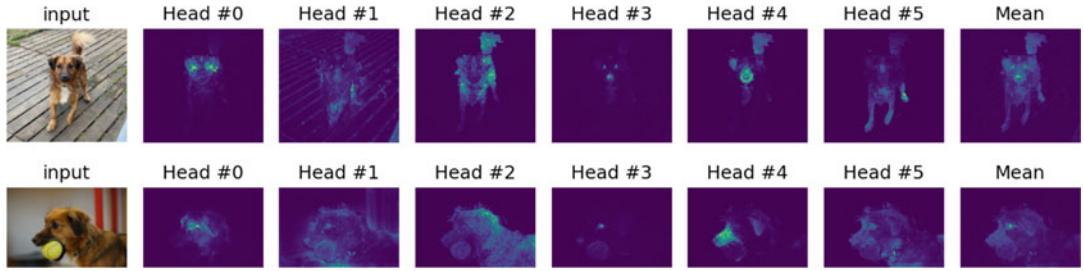


Fig. 12 DINO samples. Visualization of the attention matrix of ViT heads trained with DINO. The ViT discovers the semantic structure of an image in an unsupervised way

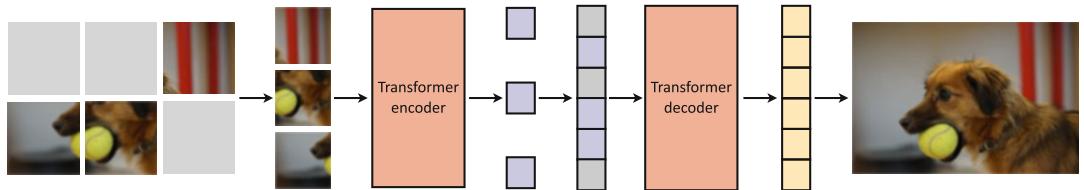


Fig. 13 The MAE training procedure. After masking some tokens of an image, the remaining tokens are fed to an encoder. Then a decoder tries to reconstruct the original image from this representation. Figure inspired from [29]

well in a given downstream task. This is illustrated in Fig. 13. One of the key observations of the MAE work [29] is that the decoder does not need to be very good for the encoder to achieve good performance: by using only a small decoder, MAE successfully trains a ViT in an autoencoder fashion.

4.4 Image Generation with Transformers and Attention

Attention and vision transformers have also helped in developing fresh ideas and creating new architectures for generative models and in particular for generative adversarial networks (GANs).

GANsformers [30] GANsformers are the most representative work of GANs with transformers, as they are a hybrid architecture using both attention and CNNs. The GANsformer architecture is illustrated in Fig. 14. The model first splits the latent vector of a GAN into multiple tokens. Then, a cross-attention mechanism is used to improve the generated feature maps, and at the same time, the GANsformer architecture retrieves information from the generated feature map to enrich the tokens. This mechanism allows the GAN to have better and richer semantic knowledge, which is showed to be useful for generating multimodal images.

StyleSwin [31] Another approach for generative modeling is to purely use a ViT architecture like StyleSwin [31]. StyleSwin is a GAN that leverages a similar type of attention as the Swin transformer [24]. This allows to generate high-definition images without having to deal with the quadratic cost problem.

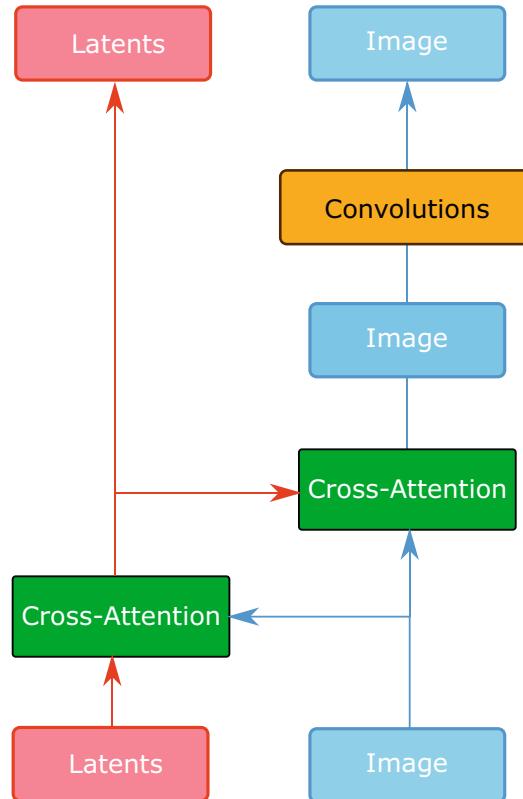


Fig. 14 GANsformer architecture. A set of latents contribute to bring information to a CNN feature map. Figure inspired from [30]

5 Vision Transformers for Other Domains

In this section, we present applications of visual transformers to other domains. First, we describe multimodal transformers operating with vision and language (Subheading 5.1), then we describe video-level attention and video transformers (Subheadings 5.2 and 5.3), and finally we present multimodal video transformers operating with vision, language, and audio (Subheading 5.4).

5.1 Multimodal Transformers: Vision and Language

As transformers have found tremendous success in both natural language processing and computer vision, their use in vision-language tasks is also of interest. In this section, we describe some representative multimodal methods for vision and language: ViLBERT (Subheading 5.1.1), DALL-E (Subheading 5.1.3), and CLIP (Subheading 5.1.2).

5.1.1 ViLBERT

Vision-and-language BERT (ViLBERT) [32] is an example of architecture that fuses two modalities. It consists of two parallel streams, each one working with one modality. The vision stream extracts

bounding boxes from images via an object detection network, by encoding their position. The language stream embeds word vectors and extracts feature vectors using the basic transformer encoder block [4] (Fig. 3 left). These two resulting feature vectors are then fused together by a cross-attention layer (Subheading 1.3.2). This follows the standard architecture of the transformer encoder block, where the keys and values of one modality are passed onto the MCA block of the other modality. The output of the cross-attention layer is passed into another transformer encoder block, and these two layers are stacked multiple times.

The language stream is initialized with BERT trained on Book Corpus [12] and Wikipedia [11], while the visual stream is initialized with Faster R-CNN [33]. On top of the pretraining of each stream, the whole architecture is pretrained on the Conceptual Captions dataset [34] on two pretext tasks.

VILBERT has been proven powerful for a variety of multimodal tasks. In the original paper, VILBERT was fine-tuned to a variety of tasks, including visual question answering, visual commonsense reasoning, referring expressions, and caption-based image retrieval.

5.1.2 CLIP

Connecting Text and Images (CLIP) [35] is designed to address two major issues of deep learning models: costly datasets and inflexibility. While most deep learning models are trained on labeled datasets, CLIP is trained on 400 million text-image pairs that are scraped from the Internet. This reduces the labor of having to manually label millions of images that are required to train powerful deep learning models. When models are trained on one specific dataset, they also tend to be difficult to extend to other applications. For instance, the accuracy of a model trained on ImageNet is generally limited to its own dataset and cannot be applied to real-world problems. To optimize training, CLIP models learn to perform a wide variety of tasks during pretraining, and this task allows for zero-shot transfer to many existing datasets. While there are still several potential improvements, this approach is competitive to supervised models that are trained on specific datasets.

CLIP Architecture and Training

CLIP is used to measure the similarity between the text input and the image generated from a latent vector. At the core of the approach is the idea of learning perception from supervision contained in natural language. Methods which work on natural language can learn passively from the supervision contained in the vast amount of text on the Internet.

Given a batch of N (image, text) pairs, CLIP is trained to predict which of the $N \times N$ possible (image, text) pairings across a batch actually occurred. To do this, CLIP learns a multimodal embedding space by jointly training an image encoder and a text encoder to maximize the cosine similarity of the image and text

embeddings of the N real pairs in the batch while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairings. A symmetric cross-entropy loss over these similarity scores is optimized.

Two different architectures were considered for the image encoder. For the first, ResNet-50 [36] is used as the base architecture for the image encoder due to its widespread adoption and proven performance. Several modifications were made to the original version of ResNet. For the second architecture, ViT is used with some minor modifications: first, adding an extra layer normalization to the combined patch and position embeddings before the transformer and, second, using a slightly different initialization scheme.

The text encoder is a standard transformer [4] (Subheading 2.1) with the architecture modifications described in [35]. As a base size, CLIP uses a 63M-parameter 12-layer 512-wide model with eight attention heads. The transformer operates on a lowercased byte pair encoding (BPE) representation of the text with a 49,152 vocab size [37]. The max sequence length is capped at 76. The text sequence is bracketed with [SOS] and [EOS] tokens,⁴ and the activations of the highest layer of the transformer at the [EOS] token are treated as the feature representation of the text which is layer normalized and then linearly projected into the multimodal embedding space.

5.1.3 DALL-E and DALL-E 2

DALL-E [38] is another example of the application of transformers in vision. It generates images from a natural language prompt—some examples include “an armchair in the shape of an avocado” and “a penguin made of watermelon.” It uses a decoder-only model, which is similar to GPT-3 [39]. DALL-E uses 12 billion parameters and is pretrained on Conceptual Captions [34] with over 3.3 million text-image pairs. DALL-E 2 [40] is the upgraded version of DALL-E, based on diffusion models and CLIP (Subheading 5.1.2), and allows better performances with more realistic and accurate generated images. In addition to producing more realistic results with a better resolution than DALL-E, DALL-E 2 is also able to edit the outputs. Indeed, with DALL-E 2, one can add or remove realistically an element in the output and can also generate different variations of the same output. These two models clearly demonstrate the powerful nature and scalability of transformers that are capable of efficiently processing a web-scale amount of data.

⁴ [SOS], start of sequence; [EOS], end of sequence

5.1.4 Flamingo

Flamingo [41] is a visual language model (VLM) tackling a wide range of multimodal tasks based on few-shot learning. This is an adaptation of large language models (LLMs) handling an extra visual modality with 80B parameters.

Flamingo consists of three main components: a vision encoder, a perceiver resampler, and a language model. First, to encode images or videos, a vision convolutional encoder [42] is pretrained in a contrastive way, using image and text pairs.⁵ Then, inspired by the perceiver architecture [25] (detailed in Subheading 1.3.2), the perceiver resampler takes a variable number of encoded visual features and outputs a fixed-length latent code. Finally, this visual latent code conditions the language model by querying language tokens through cross-attention blocks. Those cross-attention blocks are interleaved with pretrained and frozen language model blocks.

The whole model is trained using three different kinds of datasets without annotations (text with image content from web-pages [41], text and image pairs [41, 43], and text and video pairs [41]). Once the model is trained, it is fine-tuned using few-shot learning techniques to tackle specific tasks.

5.2 Video Attention

Video understanding is a long-standing problem, and despite incredible computer vision advances, obtaining the best video representation is still an active research area. Videos require employing effective spatiotemporal processing of RGB and time streams to capture long-range interactions [44, 45] while focusing on important video parts [46] with minimum computational resources [47].

Typically, video understanding benefits from 2D computer vision, by adapting 2D image processing methods to 3D spatiotemporal methods [48]. And through the Video Vision Transformer (ViViT) [49], history repeats itself. Indeed, with the rise of transformers [4] and the recent advances in image classification [5], video transformers appear as logical successors of CNNs.

However, in addition to the computationally expensive video processing, transformers also require a lot of computational resources. Thus, developing efficient spatiotemporal attention mechanisms is essential [25, 49, 50].

In this section, we first describe the general principle of video transformers (Subheading 5.2.1), and then, we detail three different attention mechanisms used for video representation (Subheadings 5.2.2, 5.2.3, and 5.2.4).

⁵The text is encoded using a pretrained BERT model [10].

5.2.1 General Principle

Generally, inputs of video transformers are RGB video clips $X \in \mathbb{R}^{F \times H \times W \times 3}$, with F frames of size $H \times W$.

To begin with, video transformers split the input video clip X into ST tokens $\mathbf{x}_i \in \mathbb{R}^K$, where S and T are, respectively, the number of tokens along the spatial and temporal dimension and K is the size of a token.

To do so, the simplest method extracts nonoverlapping 2D patches of size $P \times P$ from each frame [5], as used in TimeSformer [50]. This results in $S = HW/P^2$, $T = F$, and $K = P^2$.

However, there exist more elegant and efficient token extraction methods for videos. For instance, in ViViT [49], the authors propose to extract 3D volumes from videos (involving $T \neq F$) to capture spatiotemporal information within tokens. In TokenLearner [47], they propose a learnable token extractor to select the most important parts of the video.

Once raw tokens \mathbf{x}_i are extracted, transformer architectures aim to map them into d -dimensional embedding vectors $Z \in \mathbb{R}^{ST \times d}$ using a linear embedding $E \in \mathbb{R}^{d \times K}$:

$$Z = [\mathbf{z}_{cls}, E\mathbf{x}_1, E\mathbf{x}_2, \dots, E\mathbf{x}_{ST}] + PE, \quad (10)$$

where $\mathbf{z}_{cls} \in \mathbb{R}^d$ is a classification token that encodes information from all tokens of a single sample [10] and $PE \in \mathbb{R}^{ST \times d}$ is a positional embedding that encodes the spatiotemporal position of tokens, since the subsequent attention blocks are permutation invariant [4].

In the end, embedding vectors Z pass through a sequence of L transformer layers. A transformer layer ℓ is composed of a series of multi-head self-attention (MSA) [4], layer normalization (LN) [51], and MLP blocks:

$$\begin{aligned} \mathbf{Y}^\ell &= \text{MSA}(\text{LN}(Z^\ell)) + Z^\ell, \\ Z^{\ell+1} &= \text{MLP}(\text{LN}(\mathbf{Y}^\ell)) + \mathbf{Y}^\ell. \end{aligned} \quad (11)$$

In this way, as shown in Fig. 2, we denote four different components in a video transformer layer: the query-key-value (QKV) projection, the MSA block, the MSA projection, and the MLP. For a layer with h heads, the complexity of each component is [4]:

- *QKV projection*: $\mathcal{O}(h.(2STdd_k + STdd_v))$
- *MSA*: $\mathcal{O}(hS^2 T^2.(d_k + d_v))$
- *MSA projection*: $\mathcal{O}(SThd_v d)$
- *MLP*: $\mathcal{O}(STd^2)$

We note that the MSA complexity is the most impacting component, with a quadratic complexity with respect to the number of tokens. Hence, for comprehension and clarity purposes, in the rest of the section, we consider the global complexity of a video transformer with L layers to equal to $\mathcal{O}(LS^2 T^2)$.

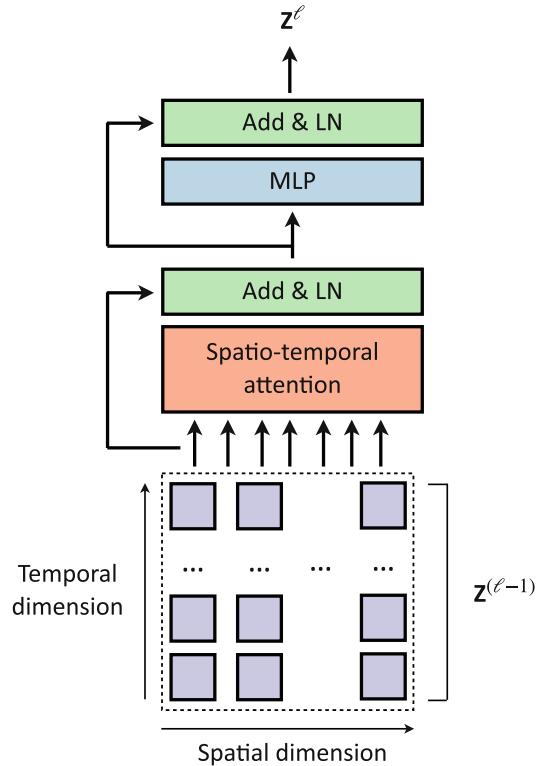


Fig. 15 Full space-time attention mechanism. Embedding tokens at layer $\ell - 1$, $\mathbf{Z}^{(\ell-1)}$ are all fed simultaneously through a unique spatiotemporal attention block. Finally, the spatiotemporal embedding is passed through an MLP and normalized to output embedding tokens of the next layer, \mathbf{z}^ℓ . Figure inspired from [50]

5.2.2 Full Space-Time Attention

As described in [49, 50], *full space-time attention* mechanism is the most basic and direct spatiotemporal attention mechanism. As shown in Fig. 15, it consists in computing self-attention across all pairs of extracted tokens.

This method results in a heavy complexity of $\mathcal{O}(LS^2T^2)$ [49, 50]. This quadratic complexity can fast be memory-consuming, in which it is especially true when considering videos. Therefore, using full space-time attention mechanism is impractical [50].

5.2.3 Divided Space-Time Attention

A smarter and more efficient way to compute spatiotemporal attention is the *divided space-time attention* mechanism, first described in [50].

As shown in Fig. 16, it relies on computing spatial and temporal attention separately in each transformer layer. Indeed, we first compute the spatial attention, i.e., self-attention within each temporal index, and then the temporal attention, i.e., self-attention across all temporal indices.

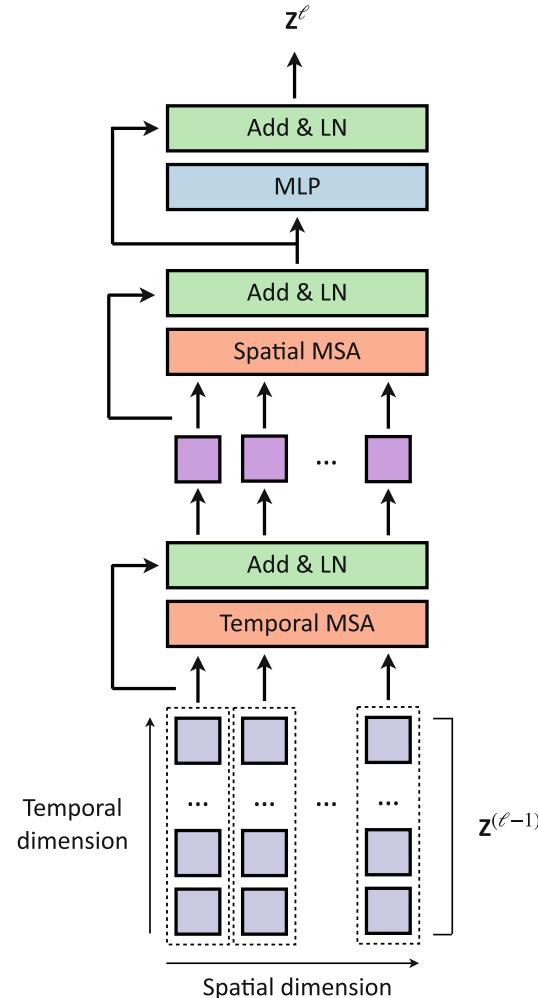


Fig. 16 Divided space-time attention mechanism. Embedding tokens at layer $\ell - 1$, $Z^{(\ell-1)}$ are first processed along the temporal dimension through a first MSA block, and the resulting tokens are processed along the spatial dimension. Finally, the spatiotemporal embedding is passed through an MLP and normalized to output embedding tokens of the next layer, Z^ℓ . Figure inspired from [50]

The complexity of this attention mechanism is $\mathcal{O}(LST.(S + T))$ [50]. By separating the calculation of the self-attention over the different dimensions, one tames the quadratic complexity of the MSA module. This mechanism highly reduces the complexity of a model with respect to the full space-time complexity. Therefore, it is reasonable to use it to process videos [50].

5.2.4 Cross-Attention Bottlenecks

An even more refined way to reduce the computational cost of attention calculation consists of using cross attention as a bottleneck. For instance, as shown in Fig. 17 and mentioned in Subheading 3.2, the perceiver [25] projects the extracted tokens x_i into a

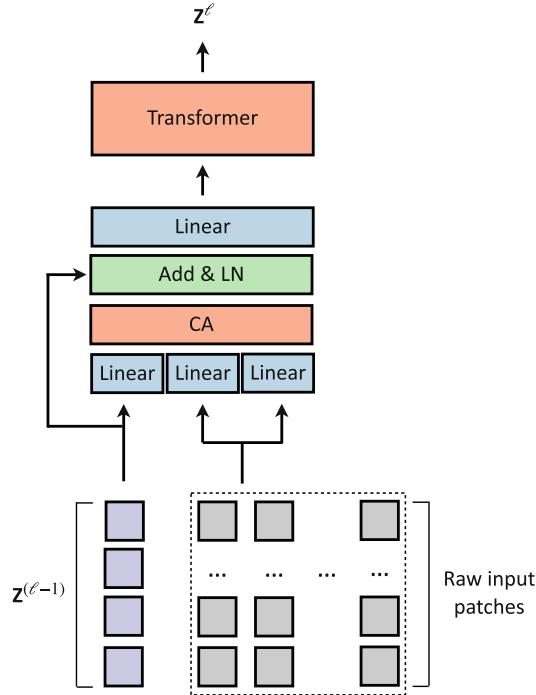


Fig. 17 Attention bottleneck mechanism. Raw input patches and embedding tokens at layer $\ell - 1$, $Z^{(\ell-1)}$ are fed to a cross-attention block (CA) and then normalized and projected. Finally, the resulting embedding is passed through a transformer to output embedding tokens of the next layer, Z^ℓ . Figure inspired from [25]

very low-dimensional embedding through a cross-attention block placed before the transformer layers.

Here, the cross-attention block placed before the L transformer layers reduce the input dimension from ST to N , where $N \ll ST$ ⁶, thus resulting in a complexity of $\mathcal{O}(STN)$. Hence, the total complexity of this attention block is $\mathcal{O}(STN + LN^2)$. It reduces again the complexity of a model with respect to the *divided space-time attention* mechanism. We note that it enables to design deep architectures, as in the perceiver [25], and then it enables the extraction of higher-level features.

5.2.5 Factorized Encoder

Lastly, the *factorized encoder* [49] architecture is the most efficient with respect to the complexity/performance trade-off.

As in *divided space-time attention*, the *factorized encoder* aims to compute spatial and temporal attention separately. Nevertheless, as shown in Fig. 18, instead of mixing spatiotemporal tokens in each transformer layer, here, there exist two separate encoders:

⁶ In practice, $N \leq 512$ for perceiver [25], against $ST = 16 \times 16 \times (32/2) = 4096$ for ViViT-L [49]

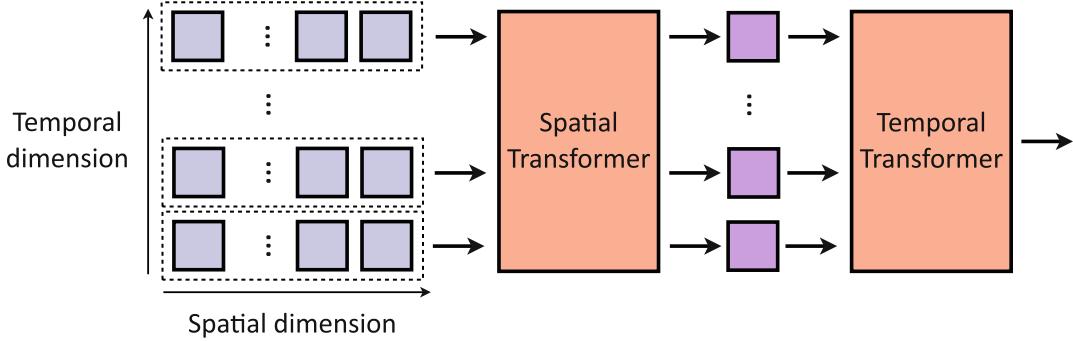


Fig. 18 Factorized encoder mechanism. First, a spatial transformer processes input tokens along the spatial dimension. Then, a temporal transformer processes the resulting spatial embedding along the temporal dimension. Figure inspired from [25]

First, a representation of each temporal index is obtained, thanks to a spatial encoder with L_s layers. Second, these tokens are passed through a temporal encoder with L_t layers (i.e., $L = L_s + L_t$).

Hence, the complexity of a such architecture has two main components: the spatial encoder complexity of $\mathcal{O}(L_s S^2)$ and the temporal encoder complexity of $\mathcal{O}(L_t T^2)$. It results in a global complexity of $\mathcal{O}(L_s S^2 + L_t T^2)$. Thus, it leads to very lightweight models. However, as it first extracts per-frame features and then aggregates them to a final representation, it corresponds to a late-fusion mechanism, which can sometimes be a drawback as it does not mix spatial and temporal information simultaneously [52].

5.3 Video Transformers

5.3.1 TimeSformer

In this section, we present two modern transformer-based architectures for video classification. We start by introducing the TimeSformer architecture in Subheading 5.3.1 and then the ViViT architecture in Subheading 5.3.2.

TimeSformer [50] is one of the first architectures with space-time attention that impacted the video classification field. It follows the same structure and principle described in Subheading 5.2.1.

First, it takes as input an RGB video clip sampled at a rate of 1/32 and decomposed into 2D 16×16 patches.

As shown in Fig. 19, the TimeSformer architecture is based on the ViT architecture (Subheading 2.3), with 12 12-headed MSA layers. However, the added value compared to the ViT is that TimeSformer uses the *divided space-time attention* mechanism (Subheading 5.2.3). Such attention mechanism enables to capture high-level spatiotemporal features while taming the complexity of the model. Moreover, the authors introduce three variants of the architecture: (i) TimeSformer, the standard version of the model, that operates on 8 frames of 224×224 ; (ii) TimeSformer-L, a configuration with high spatial resolution, that operates on 16 frames of 448×448 ; and (iii) TimeSformer-HR, a long temporal range setup, that operates on 96 frames of 224×224 .

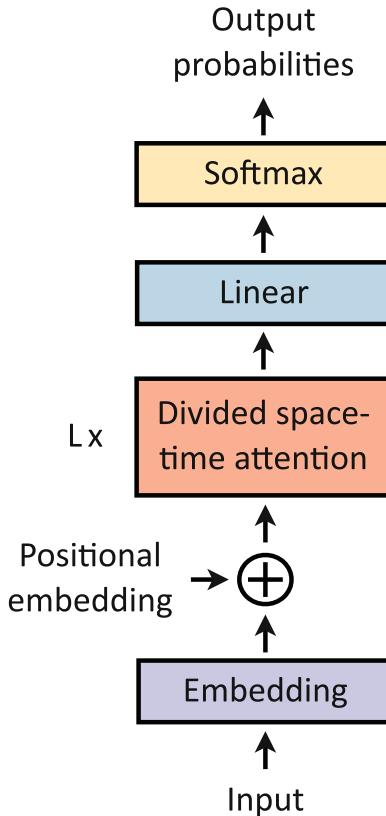


Fig. 19 TimeSformer architecture. The TimeSformer first projects input to embedding tokens, which are summed to positional embedding tokens. The resulting tokens are then passed through L divided space-time attention blocks and then linearly projected to obtain output probabilities

Finally, the terminal classification token embedding is passed through an MLP to output a probability for all video classes. During inference, the final prediction is obtained by averaging the output probabilities from three different spatial crops of the input video clip (top left, center, and bottom right).

TimeSformer achieves similar state-of-the-art performances as the 3D CNNs [53, 54] on various video classification datasets, such as Kinetics-400 and Kinetics-600 [55]. Note the TimeSformer is much faster to train (416 training hours against 3840 hours [50] for a SlowFast architecture [54]) and, also, more efficient (0.59 TFLOPs against 1.97 TFLOPs [50] for a SlowFast architecture [53]).

5.3.2 ViViT

ViViT [49] is the main extension of the ViT [5] architecture (Subheading 2.3) for video classification.

First, the authors use a 16 tubelet embedding instead of a 2D patch embedding, as mentioned in Subheading 5.2.1. This alternate embedding method aims to capture the spatiotemporal

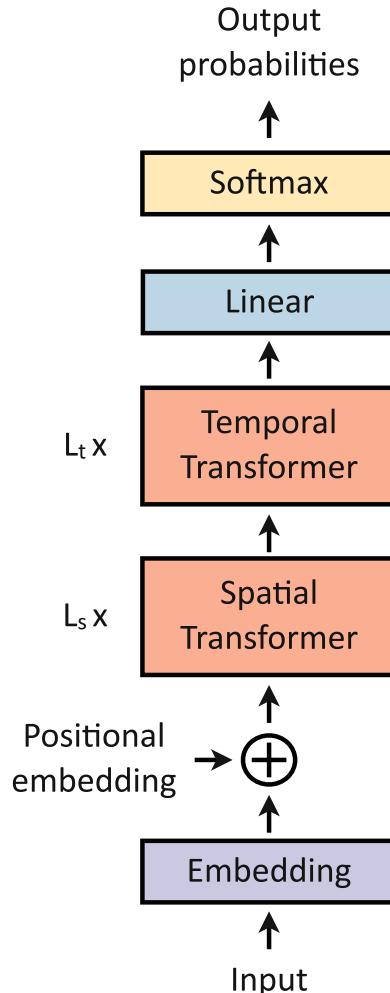


Fig. 20 ViViT architecture. The ViViT first projects input to embedding tokens, which are summed to positional embedding tokens. The resulting tokens are first passed through L_s spatial attention blocks and then through L_t temporal attention blocks. The resulting output is linearly projected to obtain output probabilities

information from the tokenization step, unlike standard architectures that fuse spatiotemporal information from the first attention block.

As shown in Fig. 20, the ViViT architecture is based on *factorized encoder* architecture (Subheading 5.2.5) and consists of one spatial and one temporal encoder operating on input clips with 32 frames of 224×224 . The spatial encoder uses one of the three ViT variants as backbone.⁷ For the temporal encoder, the number

⁷ ViT-B: 12 12-headed MSA layers; ViT-L: 24 16-headed MSA layers; and ViT-H: 32 16-headed MSA layers.

of layers does not impact much the performance, so that, according to the performance/complexity trade-off, the number MSA layers is fixed at 4. The authors show that such architecture reaches high performances while reducing drastically the complexity.

Finally, as in TimeSformer (Subheading 5.3.1), ViViT outputs probabilities for all video classes through the last classification token embedding and averages the obtained probabilities across three crops of each input clip (top left, center, and bottom right).

ViViT outperforms both 3D CNNs [53, 54] and TimeSformer [50] on the Kinetics-400 and Kinetics-600 datasets [55]. Note the complexity of this architecture is highly reduced in comparison to other state-of-the-art models. For instance, the number of FLOPs for a ViViT-L/ $16 \times 16 \times 2$ is 3.89×10^{12} against 7.14×10^{12} for a TimeSformer-L [50] and 7.14×10^{12} for a SlowFast [53] architecture.

5.4 Multimodal Video Transformers

Nowadays, one of the main gaps between artificial and human intelligence is the ability for us to process multimodal signals and to enrich the analysis by mixing the different modalities. Moreover, until recently, deep learning models have been focusing mostly on very specific visual tasks, typically based on a single modality, such as image classification [5, 17, 18, 56, 57], audio classification [25, 52, 58, 59], and machine translation [10, 60–63]. These two factors combined have pushed researchers to take up multimodal challenges.

The *default* solution for multimodal tasks consists in first creating an individual model (or network) per modality and then in fusing the resulting single-modal features together [64, 65]. Yet, this approach fails to model interactions or correlations among different modalities. However, the recent rise of attention [4, 5, 49] is promising for multimodal applications, since attention performs very well at combining multiple inputs [25, 52, 66, 67].

Here, we present two main ways of dealing with several modalities:

1. **Concatenating tokens from different modalities into one vector** [25, 66]. The multimodal video transformer (MM-ViT) [66] combines raw RGB frames, motion features, and audio spectrogram for video action recognition. To do so, the authors fuse tokens from all different modalities into a single-input embedding and pass it through transformer layers. However, a drawback of this method is that it fails to distinguish well one modality to another. To overcome this issue, the authors of the perceiver [25] propose to learn a modality embedding in addition to the positional embedding (*see* Subheadings 3.2 and 5.2.1). This allows associating each token

with its modality. Nevertheless, given that (i) the complexity of a transformer layer is quadratic with respect to the number of tokens (Subheading 5.2.1) and (ii), with this method, the number of tokens is multiplied by the number of modalities, it may lead to skyrocketing computational cost [66].

2. **Exploiting cross attention** [52, 67, 68]. Several modern approaches exploit cross attention to mix multiple modalities, such as [52] for audio and video, [67] for text and video, and [68] for audio, text, and video. The commonality among all these methods is that they exploit the intrinsic properties of cross attention by querying one modality with a key-value pair from the other one [52, 67]. This idea can be easily generalized to more than two modalities by computing cross attention across each combination of modalities [68].

6 Conclusion

Attention is an intuitive and efficient technique that enables handling local and global cues.

On this basis, the first pure attention architecture, the transformer [4], has been designed for NLP purposes. Quickly, the computer vision field has adapted the transformer architecture for image classification, by designing the first visual transformer model: the vision transformer (ViT) [5].

However, even if transformers naturally lead to high performances, the raw attention mechanism is a computationally greedy and heavy technique. For this reason, several enhanced and refined derivatives of attention mechanisms have been proposed [21–26].

Then, rapidly, a wide variety of other tasks have been conquered by transformer-based architectures, such as object detection [14], image segmentation [27], self-supervised learning [28, 29], and image generation [30, 31]. In addition, transformer-based architectures are particularly well suited to handle multidimensional tasks. This is because multimodal signals are easily combined through attention blocks, in particular vision and language cues [32, 35, 38] and spatiotemporal signals are also easily tamed, as in [25, 49, 50].

For these reasons, transformer-based architectures enabled many fields to make tremendous progresses in the last few years. In the future, transformers will need to become more and more computationally efficient, e.g., to be usable on cellphones, and will play a huge role to tackle multimodal challenges and bridge together most AI fields.

References

1. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: International conference on learning representations
2. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Empirical methods in natural language processing, association for computational linguistics, pp 1724–1734
3. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Advances in neural information processing systems, vol 27
4. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, vol 30
5. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16×16 words: transformers for image recognition at scale. In: International conference on learning representations
6. Press O, Wolf L (2017) Using the output embedding to improve language models. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers, association for computational linguistics, pp 157–163
7. Fedus W, Zoph B, Shazeer N (2021) Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. arXiv preprint arXiv:210103961
8. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 21(140):1–67
9. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M (2021) Transformers in vision: a survey. *ACM Comput Surv* 24:200
10. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), association for computational linguistics, pp 4171–4186
11. Wikimedia Foundation (2019) Wikimedia downloads. <https://dumps.wikimedia.org>
12. Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S (2015) Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the international conference on computer vision, pp 19–27
13. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
14. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: Proceedings of the European conference on computer vision. Springer, Berlin, pp 213–229
15. Ramachandran P, Parmar N, Vaswani A, Bello I, Levskaya A, Shlens J (2019) Stand-alone self-attention in vision models. In: Advances in neural information processing systems, vol 32
16. Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen LC (2020) Axial-deeplab: stand-alone axial-attention for panoptic segmentation. In: Proceedings of the European conference on computer vision. Springer, Berlin, pp 108–126
17. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 248–255
18. Krizhevsky A, Hinton G (2009) Learning multiple layers of features from tiny images. Tech rep University of Toronto, Toronto, ON
19. Sun C, Shrivastava A, Singh S, Gupta A (2017) Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the international conference on computer vision, pp 843–852
20. Selva J, Johansen AS, Escalera S, Nasrollahi K, Moeslund TB, Clapés A (2022) Video transformers: a survey. arXiv preprint arXiv:220105991
21. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H (2021) Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. PMLR, pp 10347–10357

22. d'Ascoli S, Touvron H, Leavitt ML, Morcos AS, Biroli G, Sagun L (2021) Convit: improving vision transformers with soft convolutional inductive biases. In: International conference on machine learning. PMLR, pp 2286–2296
23. Hassani A, Walton S, Shah N, Abuduweili A, Li J, Shi H (2021) Escaping the big data paradigm with compact transformers. arXiv preprint arXiv:210405704
24. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the international conference on computer vision
25. Jaegle A, Gimeno F, Brock A, Vinyals O, Zisserman A, Carreira J (2021) Perceiver: general perception with iterative attention. In: International conference on machine learning. PMLR, pp 4651–4664
26. Jaegle A, Borgeaud S, Alayrac JB, Doersch C, Ionescu C, Ding D, Koppula S, Zoran D, Brock A, Shelhamer E et al (2021) Perceiver IO: a general architecture for structured inputs & outputs. arXiv preprint arXiv:210714795
27. Strudel R, Garcia R, Laptev I, Schmid C (2021) Segmenter: transformer for semantic segmentation. In: Proceedings of the international conference on computer vision, pp 7262–7272
28. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A (2021) Emerging properties in self-supervised vision transformers. In: Proceedings of the international conference on computer vision
29. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2021) Masked autoencoders are scalable vision learners. arXiv preprint arXiv:211106377
30. Hudson DA, Zitnick L (2021) Generative adversarial transformers. In: International conference on machine learning. PMLR, pp 4487–4499
31. Zhang B, Gu S, Zhang B, Bao J, Chen D, Wen F, Wang Y, Guo B (2022) Styleswin: transformer-based GAN for high-resolution image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition
32. Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in neural information processing systems, vol 32
33. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, vol 28
34. Sharma P, Ding N, Goodman S, Soricut R (2018) Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of ACL
35. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, pp 8748–8763
36. He X, Peng Y (2017) Fine-grained image classification via combining vision and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5994–6002
37. Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers), association for computational linguistics, pp 1715–1725
38. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I (2021) Zero-shot text-to-image generation. In: International conference on machine learning. PMLR, pp 8821–8831
39. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al (2020) Language models are few-shot learners. In: Advances in neural information processing systems, vol 33, pp 1877–1901
40. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:220406125
41. Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M et al (2022) Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:220414198
42. Brock A, De S, Smith SL, Simonyan K (2021) High-performance large-scale image recognition without normalization. In: International conference on machine learning. PMLR, pp 1059–1071
43. Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, Le Q, Sung YH, Li Z, Duerig T (2021) Scaling up visual and vision-language

- representation learning with noisy text supervision. In: International conference on machine learning. PMLR, pp 4904–4916
44. Epstein D, Vondrick C (2021) Learning goals from failure. In: Proceedings of the IEEE conference on computer vision and pattern recognition
 45. Marin-Jimenez MJ, Kalogeiton V, Medina-Suarez P, Zisserman A (2019) Laco-net: revisiting people looking at each other in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition
 46. Nagrani A, Yang S, Arnab A, Jansen A, Schmid C, Sun C (2021) Attention bottlenecks for multimodal fusion. In: Advances in neural information processing systems
 47. Ryoo M, Piergiovanni A, Arnab A, Dehghani M, Angelova A (2021) Tokenlearner: Adaptive space-time tokenization for videos. In: Advances in neural information processing systems, vol 34
 48. Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and imagenet? In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6546–6555
 49. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C (2021) Vivit: a video vision transformer. In: Proceedings of the international conference on computer vision, pp 6836–6846
 50. Bertasius G, Wang H, Torresani L (2021) Is space-time attention all you need for video understanding? In: International conference on machine learning
 51. Ba JL, Kiros JR, Hinton GE (2016) Layer normalization. arXiv preprint arXiv:160706450
 52. Nagrani A, Yang S, Arnab A, Jansen A, Schmid C, Sun C (2021) Attention bottlenecks for multimodal fusion. In: Advances in neural information processing systems, vol 34
 53. Feichtenhofer C, Fan H, Malik J, He K (2019) Slowfast networks for video recognition. In: Proceedings of the international conference on computer vision, pp 6202–6211
 54. Carreira J, Zisserman A (2017) Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308
 55. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P et al (2017) The kinetics human action video dataset. arXiv preprint arXiv:170506950
 56. Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L (2021) CvT: introducing convolutions to vision transformers. In: Proceedings of the international conference on computer vision, pp 22–31
 57. Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H (2021) Going deeper with image transformers. In: Proceedings of the international conference on computer vision, pp 32–42
 58. Gemmeke JF, Ellis DP, Freedman D, Jansen A, Lawrence W, Moore RC, Plakal M, Ritter M (2017) Audio set: an ontology and human-labeled dataset for audio events. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 776–780
 59. Gong Y, Chung YA, Glass J (2021) AST: audio spectrogram transformer. In: Proceedings of interspeech 2021, pp 571–575
 60. Bojar O, Buck C, Federmann C, Haddow B, Koehn P, Leveling J, Monz C, Pecina P, Post M, Saint-Amand H, et al (2014) Findings of the 2014 workshop on statistical machine translation. In: Proceedings of the 9th workshop on statistical machine translation, pp 12–58
 61. Liu X, Duh K, Liu L, Gao J (2020) Very deep transformers for neural machine translation. arXiv preprint arXiv:200807772
 62. Edunov S, Ott M, Auli M, Grangier D (2018) Understanding back-translation at scale. In: Empirical methods in natural language processing, association for computational linguistics, pp 489–500
 63. Lin Z, Pan X, Wang M, Qiu X, Feng J, Zhou H, Li L (2020) Pre-training multilingual neural machine translation by leveraging alignment information. In: Empirical methods in natural language processing, association for computational linguistics, pp 2649–2663
 64. Owens A, Efros AA (2018) Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European conference on computer vision, pp 631–648
 65. Ramanishka V, Das A, Park DH, Venugopalan S, Hendricks LA, Rohrbach M, Saenko K (2016) Multimodal video description. In: Proceedings of the ACM international conference on multimedia, pp 1092–1096

66. Chen J, Ho CM (2022) Mm-vit: Multi-modal video transformer for compressed video action recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1910–1921
67. Narasimhan M, Rohrbach A, Darrell T (2021) Clip-it! language-guided video summarization. In: Advances in neural information processing systems, vol 34
68. Liu ZS, Courant R, Kalogeiton V (2022) FunnyNet: Audiovisual Learning of Funny Moments in Videos. In: Proceedings of the Asian conference on computer vision. Springer, Macau, China, pp 3308–3325

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third-party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part II

Data



Chapter 7

Clinical Assessment of Brain Disorders

Stéphane Epelbaum and Federica Cacciamani

Abstract

The clinical evaluation of brain diseases strictly depends on patient's complaint and observation of their behavior. The specialist, often the neurologist, chooses whether and how to assess cognition, motor system, sensory perception, and autonomic nervous system. They may also decide to request a more in-depth examination, such as neuropsychological and language assessments and imaging or laboratory tests. From the synthesis of all these results, they will be able to make a diagnosis. The neuropsychological assessment in particular is based on the collection of medical history, on the clinical observation, and on the administration of standardized cognitive tests validated in the scientific literature. It is therefore particularly useful when a neurological disease with cognitive and/or behavioral manifestation is suspected. The introduction of machine learning methods in neurology represents an important added value to the evaluation performed by the clinician to increase the diagnostic accuracy, track disease progression, and assess treatment efficacy.

Key words Clinical assessment, Neurological examination, Neuropsychology, Cognitive scores

1 Introduction

1.1 What Is a Disease? Why Are Clinical Assessments Important?

A *disease* is a specific set of processes, often biological or histological, that induce *symptoms* (subjectively felt), which negatively affect the individual's normal functioning (e.g., discomfort, pain, suffering), are often associated with a complaint, and will manifest by *signs* (objectively measured), for instance, decreased motor strength or slowed speech. Symptoms and signs taken together define a *syndrome* (e.g., headache, vomiting, stiff neck point to a meningeal syndrome), and the syndromes are contextually interpreted by physicians to hypothesize on a given disease. If, for instance, the meningeal syndrome appears brutally and is very intense, the suspected disease will be meningeal hemorrhage. If it appears subacutely over a few hours and is accompanied by a fever, the physician will rather surmise a meningitis. Box 1 introduces the main medical definitions.

A clinical evaluation is therefore requested by the patient him-/herself or by a clinician (general practitioner, specialist, psychologist, etc.). The aim is to better characterize the symptoms and the underlying disease.

Box 1 Main Medical Definitions

Disease	Physiological (biological and/or pathological) process(es) causing pejorative clinical manifestations
Symptom	Subjective manifestation of a disease (pain, memory complaint, nausea, etc.)
Sign	Objective manifestation of a disease upon medical examination (decreased reflex, elevated blood pressure, etc.)
Syndrome	Association of symptoms and signs that can be related to a set of diseases (e.g., headache, nausea, and neck stiffness are a meningeal syndrome that can correspond to either meningitis or meningeal hemorrhage)
Clinical assessment	Stereotyped interrogation, observation, and examination of an individual by a trained healthcare provider in order to collect his/her symptoms and signs to determine a syndrome and hypothesize a main disease diagnosis and differential diagnoses

During their studies, physicians learn over a few years a large quantity of diagnostic and prognostic “decision trees” based on the co-occurrence of every set of symptoms and signs. The learning is structured so that frequent and severe diseases are more studied, while rare or orphan diseases and those considered less severe are covered more briefly. For instance, the few symptoms described above will most likely be recognized and diagnosed well by any physician as well as the degree of urgency they imply. This learning is based on aggregated knowledge at one point in time which is always susceptible to change. A clear example of such changes is Alzheimer’s disease (AD) which was considered a rare form of dementia of the young from its seminal description in 1906 [1] until the 1980s when it was finally identified by numerous pathological studies to be the predominant cause of dementia in the elderly [2]. Importantly, clinical assessment requires tools to be performed, such as the famous reflex hammer used by neurologists or cognitive tests used by the neuropsychologist. Machine learning, and the decision support system that it entails, may be considered as such a tool, although it has the peculiarity of being harder to comprehend for most clinicians which may be a specific challenge for its implementation.

Every clinical assessment, whether conducted in the routine practice of medicine or in biomedical research, has to adhere to strict ethical rules that warrant the trust the patient puts in their healthcare providers. The main rules are that of beneficence; non-maleficence; respect for any individual notwithstanding their race, gender, religion, or personal beliefs; and medical confidentiality.

Finally, the current development of digital and information technologies is rapidly changing the scope of clinical assessments. Prior to consultation, auto-assessment and patient empowerment are promoted through the development of specific applications to explicitly diagnose or monitor a disease [3, 4] and patient education and access to relevant information [5]. The main issue concerning this last point is the exponential growth of these digital solutions and the risk of misinformation that can sometime lead the patient toward unethical care [6].

1.2 Peculiarities of Clinical Assessment of Brain Disorders

The brain has functionally distinct regions, so there is a topographical correspondence between the location of the lesion in the brain and the symptom. The characterization of symptoms therefore allows to trace which brain region is affected. This helps in identifying the underlying disease. The motor and sensory cortices are perfect examples of this functional topography often depicted as homunculi [7].

Clinical evaluations for brain disorders thus follow a standardized procedure. In addition to the symptoms and signs appraisal, the physician often makes an assumption as to where the nervous system is affected. This often overlaps with the syndromic definition: “frontotemporal dementia” implies that the lesions are in the frontotemporal cortices. However, this is not always the case as some diseases and syndromes still bear the name of the physician who was the first to describe it. While most neurologists know that a parkinsonism (or Parkinson syndrome) is due to basal ganglia lesions, it is not implied in its name.

2 The Neurological Examination

2.1 General Information on the Neurological Examination

The neurological examination begins with the collection of anamnestic data, that is, the complete history recalled and recounted by a patient or their entourage, including complaint, medical history, lifestyle, concurrent treatments, etc. During the collection of anamnestic data, the clinician also carefully observes patient’s behavior. The neurologist then proceeds with the examination of brain function, which is oriented by the complaint, and often includes cognitive screening tests and examination of motor system, sensitivity, and autonomic nervous system. Usually, this examination has more formal and structured parts (this can be, for example, a systematic evaluation of reflexes always in the same order or the use of a specific scale to assess sensory or cognitive function) and other

more informal ones. In fact, the clinician chooses case by case on the basis of what is required and what is available to the physician at the time of the said assessment. For example, they may use a lay journal in their office to ask a patient to describe a complex photograph in order to get a general idea of their visuospatial perception skills. This is quite time-consuming, and, depending on the patient's case, presence of entourage, and thoroughness of the clinician, an initial visit can take from 0.5 h to 2 h to capture the essential features necessary to formulate a diagnosis, prognosis, and care plan. If the neurologists deem it necessary, they may request additional tests or examinations, such as a neuropsychological evaluation, language assessment, laboratory tests, imaging tests, etc.

For applying machine learning techniques, the results of formal exams are usually more adequate because they offer quantitative measures. However, this may change over the coming years as solutions are being developed to analyze informal material. This may include clinical reports or videos of patient examinations. Another example is natural language processing tools that may help in identifying semantic deficits in patients suffering from incipient dementia [8]. The context of data acquisition is very important and can greatly impact its quality. Among the different contexts, we can cite "routine clinical practice," "retrospective or prospective observational studies," and "clinical trials" that have increasing levels of quality due to the level of standardization of data acquisition and monitoring.

2.2 Clinical Interview

A clinical interview precedes any objective assessment. It is adapted to the patient's complaint and as standardized as possible so as not to forget any question. It consists of:

- Personal and family history with, if necessary, a family tree.
- Lifestyle (including alcohol intake and smoking).
- Past or current treatments.
- As accurate as possible description of the illness made by the patient and/or their informant. It is important to know the intensity of the symptoms, their frequency, the chronological order of their appearance, the explorations already carried out, and the treatments undertaken as well as their effectiveness.

In a formal evaluation, especially in cohort studies and clinical trials, symptoms can be assessed thanks to different scales, some of which will be presented in this chapter, depending on the clinical variable of interest. These scales' results will also be used to monitor the disease evolution, notably in order to test new treatments.

The interview process is probably the most important part of the whole clinical assessment. It will allow delineating the patient's medical issue, which in turn will determine the next steps of the examination and management plan. It also creates a relation of trust that is essential for the future adhesion of the patient to the physician's propositions.

2.3 Evaluation of Cognition and Behavior

The assessment of cognition and behavior can be carried out by the neurologist using more or less in-depth tests depending on the situation, or a complete neuropsychological assessment can be requested and carried out separately by a neuropsychologist (see Subheading 3 of this chapter). The assessment of cognition is guided by the cognitive complaint of the patient and/or the informant [9]. However, on the one hand, it is possible that the patient is not fully aware of their deficits. This is a symptom called *anosognosia* (which literally means *lack of knowledge of the disease*) and is typical of various forms of dementia, including AD and frontotemporal dementia, but also brain damage due, for example, to stroke in certain regions of the brain. On the other hand, a cognitive complaint can be due to anxiety, depression, and personality traits and may have no neurological basis. The medical doctor can use simple tests in their daily practice such as the Mini-Mental State Examination (MMSE) [10]. For a more detailed description of the MMSE, please refer to Subheading 3 of this chapter.

2.4 Evaluation of Motor System

The examination of motor function starts as soon as the physician greets their patient in the waiting room. They will immediately observe the patient's walk and their bodily movements. Then, in their office, the observation will continue to search, for example, for a muscular atrophy or fascicules (i.e., muscular shudder detected by looking at the skin of the patient). This purely observational phase is followed by a formal examination, provoking objective signs.

One goal of motor assessment is to assess muscle strength. This is done segmentally, that is, carried out by evaluating the function of muscle groups that perform the same action, for example, the muscles that allow the elbow to flex. The neurologist gives a score ranging from 0 to 5, where 0 indicates that they did not detect any movement and 5 indicates normal movement strength.

A second aspect which is assessed is muscle tone. It is explored by passively mobilizing the patient joints. Hypertonia, or rigidity, is an increase in the tone. When the neurologist moves the joint, it may remain rigidly in that position (*plastic* or *parkinsonian* hypertonia), or the limb may immediately return to the resting position as soon as the neurologist stops manipulating it (*spastic* or *elastic* hypertonia). Hypotonia is a reduction of muscle tone, i.e., lack of tension or resistance to passive movement. This is observed in cerebellar lesions and chorea.

Another goal of motor assessment is evaluating deep tendon reflexes. Using a reflex hammer, the neurologist taps the tendons (e.g., Achilles' tendon for the Achillean reflex). The deep tendon reflexes will be categorized as (1) normal, (2) increased and polykinetic (i.e., a single tap provokes more than one movement), (3) diminished or abolished (as in peripheral nervous system diseases), and (4) pendular (as in cerebellar syndrome). Often evidenced in case of increased reflexes, Babinski's sign is the lazy and

majestic extension of the big toe followed by the other toes in response to the scraping of the outer part of the foot plant. It is pathognomonic (i.e., totally specific) of a pyramidal syndrome, which is named after the axonal fiber tract that is altered: the pyramidal fasciculus. Motor assessment also includes evaluation of tremors and posture.

Once again, specific scales exist to robustly and homogeneously assess some of these signs such as the Unified Parkinson's disease rating scale (UPDRS) in Parkinson's disease [11]. For more information, the Movement Disorder Society UPDRS Revision Task Force has made the questionnaire available [12]. We report MDS-UPDRS items in Box 2. There are 65 items, 60 of which with a score from 0 to 4 (0, normal; 1, slight; 2, mild; 3, moderate; and 4, severe) and 5 with yes/no responses.

Box 2 MDS-UPDRS Structures

Part I: Non-motor experiences of daily living 13 items. Less than 10 min 1. Cognitive impairment 2. Hallucinations and psychosis 3. Depressed mood 4. Anxious mood 5. Apathy 6. Features of dopamine dysregulation syndrome 7. Nighttime sleep problems 8. Daytime sleepiness 9. Pain and other sensations 10. Urinary problems 11. Constipation problems 12. Lightheadedness on standing 13. Fatigue	Part II: Motor experiences of daily living 13 items. It does not involve examiner time; items are answered by the patient or caregiver independently. 1. Speech 2. Salivation and drooling 3. Chewing and swallowing 4. Eating tasks 5. Dressing 6. Hygiene 7. Handwriting 8. Doing hobbies and other activities 9. Turning in bed 10. Tremor 11. Getting out of bed, car, or deep chair 12. Walking and balance 13. Freezing
Part III: Motor examination 33 items (18 items with different duplicates corresponding to the right or left side or to different body parts). 15 min 1. Speech 2. Facial expression 3. Rigidity of neck and four extremities 4. Finger taps 5. Hand movements	Part IV: Motor complications Six items. 5 min 1. Time spent with dyskinesia 2. Functional impact of dyskinesias 3. Time spent in the OFF state 4. Functional impact of fluctuations 5. Complexity of motor fluctuations 6. Painful OFF-state dystonia

(continued)

Box 2 (continued)

6. Pronation/supination
7. Toe tapping
8. Leg agility
9. Arising from chair
10. Gait
11. Freezing of gait
12. Postural stability
13. Posture
14. Global spontaneity of movement
15. Postural tremor of hands
16. Kinetic tremor of hands
17. Rest tremor amplitude
18. Constancy of rest tremor

2.5 Evaluation of Sensitivity

Sensitivity is the ability to feel different tactile sensations: normal (or crude) tact, pain, hot, or cold. Once again, it depends on the anatomical regions and tracts affected by a pathological process. The anterior spinothalamic tract carries information about crude touch. The lateral spinothalamic tract conveys pain and temperature. Assessment includes measuring:

- Epicritic sensitivity: test the patient's ability to discriminate two very close stimuli.
- Deep sensitivity: test the direction of position of the joints by the blind prehension. The doctor can also ask the patient if the vibrations of a diapason on joint bones (knee, elbow) are felt.
- Discrimination of hot and cold; sensitivity to pain.

2.6 Other Evaluations

The physician evaluation will also assess the autonomic nervous system which, when impaired, can induce tensile disorders: hypo-/hypertension, orthostatic hypotension (without compensatory acceleration of pulse), diarrhea, sweating disorders, accommodation disorders, and sexual disorders. They will also evaluate cerebellar functions: balance, coordination (which when impaired causes ataxia), and tremor.

Finally, clinicians will assess cranial nerves' functions. Cranial nerves are those coming out of the brainstem and have various functions including olfaction, vision, eye movements, face sensorimotoricity, and swallowing. They are tested once again in a standar-dized way from the first one to the twelfth.

2.7 Summary of the Neurological Evaluation

At the end of this examination, the signs and symptoms are described in the report, and the physician specifies:

- A syndromic group of signs and symptoms
- The presumed location of brain damage

- A main diagnostic hypothesis
- Possibly, secondary hypotheses (differential diagnosis)
- Additional examination strategy through neuroimaging or additional examinations to refine disease diagnosis
- A therapeutic program

3 Neuropsychological Assessment

3.1 Generalities on Neuropsychological Assessment

Neuropsychology is concerned with how cognitive functions (*see* Box 3) and behavior are correlated with anatomo-physiological brain mechanisms. Thanks to the scientific-technological advances made in recent decades and the advent of increasingly sensitive structural and functional imaging techniques, we have discovered that human cognition has a modular architecture in which each module—whose operationalization depends on the reference framework—corresponds to a specific function [13]. This allowed us to understand which brain regions or structures we expect to be damaged when we observe a certain cognitive deficit [14–17]. The role of the neuropsychologist can be summarized in two core activities: assessment and intervention. In this chapter, we will focus on neuropsychological assessment, which produces data that is typically used by machine learning algorithms.

Neuropsychological assessment includes a clinical interview, followed by the measurement of cognitive functions using standardized tests and finally the interpretation of the results. This is applicable in diagnostic settings, to monitor disease progression if the diagnosis has previously been made or to measure the effectiveness of a treatment.

Box 3 Main Cognitive Functions

Memory	Short-term memory or working memory temporarily retains few pieces of information for the time needed to perform a certain task, using mechanisms such as mental repetition Episodic memory allows long-term conscious memory of a potentially infinite number of events (episodes) and contexts (time and place) in which they occurred Semantic memory allows the long-term conscious memory of a potentially infinite number of facts, concepts, and vocabulary, which constitute the knowledge that the individual has of the world Procedural memory is the memory of how things are done (e.g., tying shoelaces) and how objects are used
--------	--

(continued)

Box 3 (continued)

Attention	Selective attention is the ability to select relevant information from the environment Sustained attention is the ability to persist for a relatively long time on a certain task
Visuospatial abilities	Estimation of spatial relationships between the individual and the objects and between the objects themselves and identification of visual characteristics of a stimulus such as its orientation
Language	Oral and written production and comprehension, at a phonological, morphological, syntactic, semantic, and pragmatic level
Executive functions	Superior cognitive functions such as planning, organization, performance monitoring, decision-making, mental flexibility, etc.
Social cognition	Using information previously learned more or less explicitly to explain and predict one's own behavior and that of others in social situations

Neuropsychology is therefore an interdisciplinary discipline. It is first and foremost a branch of psychology. The clinical interview that precedes the administration of tests is typical of psychological disciplines. The clinician collects anamnestic information (i.e., regarding medical history, lifestyle, and familiarity), observes patient behavior, and builds a relationship of trust and collaboration with him/her. All of these are crucial aspects in any type of psychological interview. In addition, the neuropsychologist must also be able to understand whether the cognitive complaint or the deficits detected are linked to brain damage or whether they are psychogenic. To do this, they assess, qualitatively or quantitatively depending on the situation, the mood of the patient and the presence of any anxiety syndromes, psychotic symptoms, etc.

Neuropsychology also has obvious points in common with neurology, since it is interested in the evaluation and intervention on the cognitive-behavioral manifestation of pathologies of the central nervous system. Over the past decades, much knowledge has been gained on the relationship between cognition and brain, and many tests have been developed. As a result, neuropsychological assessment has split off from neurological examination, assuming a separate role [18].

3.2 Psychometric Properties of Neuropsychological Tests

The use of cognitive tests is the specificity of the neuropsychological assessment.

Each new test is developed according to a rigid and rigorous methodology, trying to minimize all possible sources of error or bias, and based on scientific evidence. For example, a test that aims to assess learning skills might include a list of words for the participant to memorize and then recall. These words will not be randomly selected but carefully chosen based on characteristics such as frequency of use, length, phonology, etc. The procedures for administering neuropsychological tests are also standardized. The situation (i.e., materials, instructions, test conditions, etc.) is the same for all individuals and dictated by the administration manuals provided with each test.

All tests, before being published, are validated for their psychometric properties and normed. A normative sample is selected according to certain criteria which may change depending on the situation [19]. In most cases, these are large samples of healthy individuals from the general population, stratified by age, sex, and/or level of education. In other cases, more specific samples are preferred. The goal is to identify how the score is distributed in the normative sample. In this way, we can determine if the score obtained by a hypothetical patient is normal (i.e., around the average of the normative distribution) or pathological (i.e., far from the average). Establishing how far from the average an observation must be in order to be considered abnormal is a real matter of debate [20]. Many neuropsychological scores, as well as many biological or physical attributes, follow a normal distribution in the general population. The most used metrics to determine pathology thresholds are z scores and percentiles. For a given patient, the neuropsychologist usually computes the z score by subtracting the mean of the normative sample from the raw score obtained by the patient and then dividing the result by the standard deviation (SD) of the normative sample. The distribution of z scores will have a mean of 0 and a SD of 1. We can also easily find the percentile corresponding to the z score. Most often, a score below the fifth percentile (or z score = -1.65) or the second percentile (or z -score = -2) is considered pathological. As an example, intelligence, or intelligence quotient (IQ), is an attribute that follows a normal distribution. It is conventionally measured with the Wechsler Adult Intelligence Scale, also known as WAIS [21], or the Wechsler Intelligence Scale for Children, also known as WISC [22]. The distribution of IQs has a mean of 100 and a SD of 15 points. Around 68% of individuals in the general population achieve an IQ of 100 ± 15 points. Scores between 85 and 115 are therefore considered to be average IQs (therefore normal). Ninety-five percent of individuals are in a range within 30 points of 100, thus

between 70 and 130. Scores between 70 and 85 and those between 115 and 130 indicate borderline intelligence and medium-to-higher intelligence, respectively. Finally, only a little more than 2% of people are located in the two tails, respectively. An IQ below 70 is therefore considered pathological and indicative of intellectual disability. An IQ above 130 is indicative of superior intelligence.

Another reason a new test is administered to a normative sample is to evaluate its psychometric properties to understand whether it is suitable for clinical or research use [23]. The two main properties worth mentioning are reliability and validity [24].

Reliability indicates the consistency of a measure or in other words the proportion of variance in the observed scores attributable to the actual variance of the measured function, and not to measurement errors [25]. Reliability may be assessed in various ways. Internal consistency, for example, indicates whether the items of a test all measure the same cognitive function. A common procedure to evaluate it is to randomly divide the test into two halves and calculate the correlation between them. Test-retest reliability indicates the ability of a test to provide the same score consistently over time. No undesirable event, such as a pathological event, should have occurred between the two assessments and cause the patient to score worse (or better) on the second one. Another bias that could undermine test-retest reliability is practice effect, which refers to a gain in scores that occurs when the respondent is retested with the same cognitive test. This gain does not reflect a real improvement in the function assessed [26]. Parallel forms of the same test are often used to avoid these problems. Another measure of reliability is the consistency between different examiners (inter-rater reliability). In fact, despite the standardization described above, some degree of variance may remain between examiners [27].

Validity is the capacity of a test to measure what it actually proposes to measure and not similar constructs [28]. The validity of a test can be assessed by calculating the correlation between the score of interest with another measure that is theoretically supposed to be correlated. The following are some types of validity commonly assessed when developing or validating a new neuropsychological test: content validity (i.e., the test only measures what it is supposed to measure), substantive validity (i.e., the test is developed on the basis of theoretical knowledge and empirical evidence), convergent validity (i.e., individuals belonging to a certain homogeneous group have a similar score on the same test), and divergent validity (i.e., individuals belonging to two different groups have different scores on the same test, e.g., patients versus controls).

3.3 Realization of a Neuropsychological Assessment and Interpretation of Its Results

During an assessment, the neuropsychologist chooses the most appropriate tests for the patient, ensures that they are performed correctly, and interprets their results. Indeed, each neuropsychological assessment is tailored to the patient's needs. To assess a certain cognitive function, the clinician can choose a specific test depending on the patient's level of education, the presence of any sensory deficits (e.g., tests involving verbal material will be proposed to a visually impaired patient), as well as the diagnostic hypothesis.

Once anamnestic data has been collected and the cognitive scores have been obtained, the goal is to interpret these results and define the patient's cognitive profile. Defining a cognitive profile means identifying which cognitive functions are preserved and which are impaired. In the event that one or more impaired cognitive deficits are detected, it is necessary to specify at what level the deficit is located and its severity. For example, a patient may have a memory disorder whose severity can be identified by comparing their score to normative data as described above. Depending on the test used, the neuropsychologist will be able to define whether this memory disorder is due to difficulties in creating new memory traces (linked to the medial temporal lobe [14]), or to difficulties in retrieving existing traces (linked to the prefrontal lobe [16]), and so on. By describing the impaired and preserved cognitive mechanisms and by referring to what we know about brain correlates of cognitive function, the neuropsychologist will be able to detect a pattern. This may be a cortical syndrome, such as in the event of alteration of language or visuospatial functions [29]; a subcortico-frontal profile, involving, for example, impaired executive functions [30]; a subcortical profile, often involving slow information processing [31]; etc.

It is important to clarify that the aim of the neuropsychological assessment is not to diagnose a disease, but to describe a cognitive profile. This is only one of the elements taken into account by a physician, often a neurologist, to make the diagnosis. The physician will determine which disease or pathological condition underlies the cognitive impairment, by combining the evidence from other tests, such as laboratory tests, imaging, and neurological examination, as described above.

3.4 The Example of a Cognitive Test: The Mini-Mental State Examination (MMSE)

The Mini-Mental State Examination, also known as MMSE, is one of the most widely used tools in both clinical practice and research, validated in many languages and adapted to administration in many countries. It is a screening tool for adults, which allows assessing global cognition quickly and easily through a paper-pencil test lasting 5–10 min.

Box 4 MMSE Questions and Scoring System**Temporal orientation** [5 points, 1 per item]

The respondent is asked to say the day of the week, the day of the month, the month, the year, and the season

Spatial orientation [5 points, 1 per item]

The respondent is asked to say the floor and the name of the hospital or practice, district, town, and country.

Short-term memory [3 points, 1 per word]

The examiner names three objects (apple, table, and penny in the English version), and the respondent repeats them immediately

Attention [5 points, 1 per subtraction]

The respondent subtracts 7 from 100 five times

Verbal learning [3 points, 1 per word]

The respondent recalls the three previously learned words

Denomination [2 points, 1 per object]

The respondent names two objects indicated by the examiner, often a pen and a watch

Repetition [1 point]

The respondent repeats the sentence “No ifs, ands, or buts”

Listening comprehension [3 points, 1 per task]

The respondent is asked to take a sheet with their right hand, fold it in half, and throw it on the ground

Written comprehension [1 point]

The respondent executes a written command, often “Close your eyes”

Writing [1 point]

The respondent writes a sentence that contains a verb and a subject

Praxico-constructive and visuospatial skills [1 point]

Copy of two intersecting pentagons showed by the examiner

The MMSE includes 30 questions, each with a binary score (0 for wrong answer and 1 for correct answer). More details are presented in Box 4. The total score ranges from 0 to 30. An MMSE score of 18 or less indicates severe impairment of cognitive functions. A score between 18 and 24 indicates moderate to mild impairment. A score of 25 is considered borderline. And a score of 26–30 indicates cognitive normality. Different diagnostic thresholds have been proposed as they depend—mainly—on age, education, and setting [32]. In clinical settings, a score below 24 is commonly considered pathological [33]. In research contexts, it is more common to use a cut-off of 26 (pathological if <26) [34]. The MMSE is therefore very useful for getting an idea of the patient’s cognitive functioning, also facilitating effective communication between professionals.

Concerning psychometric properties, internal consistency is reported to vary significantly according to the setting. Alpha coefficient was around 0.30 in the general population [35] and 0.96 in a clinical setting [36]. Lower coefficients may be related to lower variability in community-based samples where the majority of participants are healthy and often highly educated. Regarding test-retest reliability, healthy individuals scored better at retest (about one point higher) when they repeated the MMSE about 3 months after the first assessment. Patients with cognitive impairment, on the contrary, did not show such learning. In [10], the MMSE also had good validity in discriminating patients with Alzheimer's dementia, depression, and schizophrenia.

4 Clinical Examination by Pathology

Neurology is a broad branch of medicine that deals with all pathologies affecting the central and peripheral nervous system, also including blood vessels and muscles, such as neurodegenerative diseases, epilepsy, sleep disorders, vascular diseases, headaches, movement disorders, neuro-oncology, etc. Clinical evaluation is therefore tailored to the complaint and symptoms. The purpose is to propose a treatment or follow the evolution of the disease. There is therefore a need for sensitive clinical tests that allow for early detection of abnormalities, so that treatment can be administered more promptly.

4.1 Diversity of Brain Disorders and Clinical Evaluation

As science advances, medicine is getting increasingly specialized. Although “general neurologists” are the majority in the domain, the field is segmented in different subspecialties in university hospitals, each with their topic and diseases of interest, and dedicated tools for innovative studies. We briefly describe these subspecialties below (*see Box 5*).

Box 5 Non-exhaustive List of the Main Neurological Diseases

Neurodegenerative disorders affecting mostly cognition or behavior	Alzheimer's disease Frontotemporal dementia Lewy body dementia Primary progressive aphasia
Movement disorders	Parkinson's disease Essential tremor Dystonia

(continued)

Epilepsy	Generalized idiopathic epilepsy Absence Partial idiopathic epilepsy Secondary epilepsy (post-traumatic, post-stroke, etc.)
Stroke or neurovascular diseases	Ischemic stroke Brain hemorrhage Cerebral venous thrombosis
Neuro-oncology	Meningioma Oligodendroglioma Astrocytoma Glioblastoma Brain metastasis
Peripheral nerve diseases	Mononeuropathy Polyneuropathy Radiculopathy Plexopathy
Headaches	Migraine Tension-type headache
Sleep disorders	Sleep apnea Narcolepsy
Inflammatory and demyelinating brain diseases	Multiple sclerosis Sarcoidosis
Neurogenetic diseases	Huntington's chorea Spinocerebellar ataxia
Neuromuscular disorders	Amyotrophic lateral sclerosis Myasthenia Myopathies

4.1.1 Neurodegenerative Disorders Affecting Mostly Cognition or Behavior

They include Alzheimer's disease, Lewy body and frontotemporal dementias, as well as rarer conditions such as primary progressive aphasia. This field relies heavily on neuropsychological evaluation. Although progress has been achieved in diagnosis of these conditions (especially Alzheimer's disease) these last decades, therapeutic unmet needs remain high.

4.1.2 Movement Disorders

These include Parkinson's disease but also dystonia, myoclonus, tics, and tremors. Different treatment options have emerged for this group of diseases in the last years. These include drugs based on the dopamine levels in the brain (one of the main neurotransmitters for movement) and deep brain stimulation which requires the implantation of electrodes to stimulate or inhibit specific regions of the basal ganglia.

4.1.3 Epilepsy

This broad term refers to the abnormal electric activity of neurons in brain regions or in the whole brain inducing seizures. They are defined by the co-occurrence of symptoms or signs, and these electric abnormalities are detected by electroencephalography (EEG). Many anti-epileptic drugs exist to decrease the seizure frequency in these patients. Some patients present with pharmacoresistant epilepsy. For such patients, surgery, which aims at resecting part of the brain in order to suppress seizures, can be a treatment option.

4.1.4 Stroke or Neurovascular Diseases

Acute stroke is managed in stroke emergency units. A stroke can be either a brain infarction or a hemorrhage. They are not primary diseases of the brain tissue but of the arteries, capillaries, and veins that irrigate it. Treatment options range from rapid clot removal in ischemia (whether by thrombolysis or neuroradiological intervention), anti-aggregating or anticoagulation therapy, and physical or speech rehabilitation.

4.1.5 Neuro-oncology

This specialty deals with brain tumors, which may be malignant or benign. There are close connections with neurosurgery units and neuropathology which play a valuable role in analyzing the microstructure of the tumor in order to achieve a precise diagnosis. Treatments typically rely on a combination of surgery, radiotherapy, and chemotherapy.

4.1.6 Peripheral Nerve Diseases

They include all the diseases of the nerves outside of the brain, brainstem, or spine. These diseases induce motor, sensory, and autonomous impairments and are diagnosed through a combination of medical examination and electromyographic (EMG) recordings. Treatment options are very dependent on the cause of the disease which can range from simple mechanic compression of a nerve requiring mild surgery (carpal syndrome) to hepatic graft in some rare conditions (TTR mutation causing familial transthyretin amyloidosis).

4.1.7 Headaches

Although headaches are highly prevalent, specialists are rare in university hospital as these conditions (including migraine) are often cared for in private practice offices, except for the most urgent causes which are managed by emergency units. Treatments aim to decrease the frequency of the crisis (preventative treatments) for the most severe cases or the pain during a given crisis.

4.1.8 Sleep Disorders

Sleep disorders are sometimes managed by neurologists for some diseases (like narcolepsy) or pneumologists (since sleep apneas are among the most frequent cause of sleep impairment) or psychiatrists (tackling insomnia, often associated with psychiatric comorbidities). A sleep recording called polysomnography is sometimes

required to assess the most complex problems. Physicians can prescribe continuous positive airway pressure devices which keep the airways opened during sleep.

4.1.9 Inflammatory and Demyelinating Brain Diseases

The most emblematic of this group is multiple sclerosis in which the autoimmune system turns against the individual, penetrates the blood–brain barrier, and attacks the myelin which allows the rapid diffusion of the neuronal electric signal along the axons. This is one of the most advanced fields of neurology regarding treatment. Since the start of the twenty-first century, specific therapies preventing the crossing of the blood–brain barrier of lymphocytes revolutionized the management of multiple sclerosis [37].

4.1.10 Neurogenetic Diseases

Neurogenetic diseases are a group of rare diseases (like Huntington's chorea) due to a genetic mutation. These diseases usually follow a Mendelian mode of inheritance. They have the particularity to be detectable (through genetic testing after a specific counseling) which gives the opportunity to study them in their premorbid phase (i.e., before the onset of typical symptoms in a group of mutation carriers). Innovative gene therapies are actually being developed in some of these neurogenetic conditions [38]. Note that there also exist genetic forms of diseases which are in majority sporadic (e.g., familial forms of Alzheimer's disease).

4.1.11 Neuromuscular Disorders

These are diseases affecting the motor neurons such as amyotrophic lateral sclerosis, the neuromuscular synapse like myasthenia, or specifically the muscles in myopathies. To the exception of myasthenia, few treatment options exist in this particular field of neurology.

4.2 Importance of a Correct and Timely Diagnostic Classification

Neurologists have a saying: "time is brain." The correct and timely identification of a neurological disease is indeed crucial to be able to mitigate and sometimes reverse the signs and symptoms. As such, machine learning techniques may be very useful tools both in the context of slow-paced diseases such as Alzheimer's which are often diagnosed quite late or not at all [39] and to optimize the patient flow in emergency care, in case of stroke, for instance. This framework is theoretical as in practice some diseases can interact to induce symptoms. For instance, dementia is often of mixed origin, due to the association of degenerative (Alzheimer's disease) and vascular alterations. A walking deficit can be due to Parkinson's disease but also in part to arthrosis, etc. The correct identification of a disease is in part probabilistic, and this can lead to heterogeneity in the collected data from the clinical assessment.

5 Conclusion

Clinical assessment is central in neurology for the assessment of the patient because it is the direct reflection of what he/she feels and experiences. Indeed, according to regulatory agencies, a treatment is deemed effective if it has an effect on the clinical expression of the disease (e.g., on cognition, motor skills, sensitivity, autonomy, and survival) and not on intermediate markers such as imaging, biology, or others.

Machine learning is bringing clinical evaluation into a new era because it allows to go beyond the intuitions of the individual physician and could associate signs that were previously not seen as part of a disease type or subtype. However, the researcher should always remember that the best algorithm is only as good as the data it runs on, which depends on the clinician's understanding of how and why these particular data are collected and will be used for. So, for discovery, validation, and clinical implementation of new machine learning techniques, basic knowledge of the possible discrepancies and biases one may experience going from research setting to clinical practice is paramount.

Acknowledgments

The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6).

References

1. Lage JMM (2006) 100 years of Alzheimer's disease (1906–2006). *J Alzheimers Dis* 9(s3): 15–26
2. McKhann G, Drachman D, Folstein M et al (1984) Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 34(7): 939–944. <https://doi.org/10.1212/wnl.34.7.939>. [doi]
3. Öhman F, Hassenstab J, Berron D et al (2021) Current advances in digital cognitive assessment for preclinical Alzheimer's disease. *Alzheimers Dement* 13(1):e12217
4. Dillenseger A, Weidemann ML, Trentzsch K et al (2021) Digital biomarkers in multiple sclerosis. *Brain Sci* 11(11):1519
5. Rodríguez-Gómez O, Rodrigo A, Iradier F et al (2019) The MOPEAD project: advancing patient engagement for the detection of "hidden" undiagnosed cases of Alzheimer's disease in the community. *Alzheimers Dement* 15(6):828–839
6. Daly T, Mastroleo I, Gorski D et al (2020) The ethics of innovation for Alzheimer's disease: the risk of overstating evidence for metabolic enhancement protocols. *Theor Med Bioeth* 41(5–6):223–237
7. Penfield W, Rasmussen T (1950) The cerebral cortex of man; a clinical study of localization of function. Macmillan
8. Tsai AC, Hong SY, Yao LH et al (2021) An efficient context-aware screening system for Alzheimer's disease based on neuropsychology test. *Sci Rep* 11(1):18570

9. Cacciamani F, Houot M, Gagliardi G et al (2021) Awareness of cognitive decline in patients with Alzheimer's disease: a systematic review and meta-analysis. *Front Aging Neurosci* 13
10. Folstein MF, Folstein SE, McHugh PR (1975) "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 12(3):189–198. [https://doi.org/10.1016/0022-3956\(75\)90026-6](https://doi.org/10.1016/0022-3956(75)90026-6). [pii]
11. Ramaker C, Marinus J, Stiggebout AM et al (2022) Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. *Mov Disord* 17(5):867–876
12. Goetz CG, Tilley BC, Shaftman SR et al (2008) Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov Disord* 23(15):2129–2170
13. Vallar G (2000) The methodological foundations of human neuropsychology: studies in brain-damaged patients. In: Boller F, Grafman J (eds) *Handbook of neuropsychology*, 2nd edn. Elsevier, Amsterdam, pp 459–502
14. Preston AR, Eichenbaum H (2013) Interplay of hippocampus and prefrontal cortex in memory. *Curr Biol* 23(17):R764–R773
15. Indefrey P, Levelt WJM (2000) The neural correlates of language production. In: Gazzaniga MS (ed) *Anonymous the new cognitive neurosciences*, 2nd edn. MIT Press, Cambridge, MA, pp 845–865
16. Robinson H, Calamia M, Gläscher J et al (2014) Neuroanatomical correlates of executive functions: a neuropsychological approach using the EXAMINER battery. *J Int Neuropsychol Soc* 20(1):52–63
17. Goodale MA, Milner AD (1992) Separate visual pathways for perception and action. *Trends Neurosci* 15(1):20–25
18. Benton AL (1988) Neuropsychology: past, present and future. In: Boller F, Grafman J (eds) *Anonymous handbook of neuropsychology*. Amsterdam, pp 3–27
19. Schmidt S, Pardo Y (2014) Normative data. In: Michalos AC (ed) *Encyclopedia of quality of life and well-being research*. Springer, Dordrecht
20. Guilmette TJ, Sweet JJ, Hebben N et al (2020) American Academy of Clinical Neuropsychology consensus conference statement on uniform labeling of performance test scores. *Clin Neuropsychol* 34(3):437–453
21. Wechsler D (2008) *Wechsler adult intelligence scale—fourth edition administration and scoring manual*. Pearson, San Antonio, TX
22. Wechsler D (2003) *The Wechsler intelligence scale for children—fourth edition*. Pearson, London
23. Portney LG, Watkins MP (2009) *Foundations of clinical research: applications to practice*. Pearson Education, New Jersey
24. Anastasi A, Urbina S (1997) *Psychological testing*. Prentice Hall, Pearson Education
25. Nunnally JC, Bernstein IH (1994) *Psychometric theory*. McGraw-Hill, New York
26. Cacciamani F, Salvadori N, Eusebi P et al (2017) Evidence of practice effect in CANTAB spatial working memory test in a cohort of patients with mild cognitive impairment. *Appl Neuropsychol Adult* 22:1–12. <https://doi.org/10.1080/23279095.2017.1286346>
27. Hallgren KA (2012) Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quantit Methods Psychol* 8(1):23–34
28. Messick S (1989) Validity. In: Linn RL (ed) *Educational measurement*. American Council on education and Macmillan, New York, NY, pp 13–104
29. Huber SJ, Shuttleworth EC, Paulson GW et al (1986) Cortical vs subcortical dementia: neuropsychological differences. *Arch Neurol* 43(4):392–394
30. Bonelli RM, Cummings JL (2007) Frontal-subcortical circuitry and behavior. *Dialogues Clin Neurosci* 9(2):141–151
31. Mayeux R, Stern Y (1987) Subcortical dementia. *Arch Neurol* 44(2):129–131
32. Arevalo-Rodriguez I, Smailagic N, Roqué Figuls M et al (2015) Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database Syst Rev* 3
33. Crum RM, Anthony JC, Bassett SS et al (1993) Population-based norms for the Mini-Mental State Examination by age and educational level. *JAMA* 269(18):2386–2391
34. Meiran N, Stuss DT, Guzman DA et al (1996) Diagnosis of dementia. Methods for interpretation of scores of 5 neuropsychological tests. *Arch Neurol* 53(10):1043–1054
35. Hopp GA, Dixon RA, Grut M et al (1997) Longitudinal and psychometric profiles of two cognitive status tests in very old adults. *J Clin Psychol* 53(7):673–686

36. Foreman MD (1987) Reliability and validity of mental status questionnaires in elderly hospitalized patients. *Nurs Res* 36(4):216–220
37. Polman CH, Uitdehaag BM (2003) New and emerging treatment options for multiple sclerosis. *Lancet Neurol* 2(9):563–566
38. Hinderer C, Miller R, Dyer C et al (2020) Adeno-associated virus serotype 1-based gene therapy for FTD caused by GRN mutations. *Ann Clin Transl Neurol* 7(10):1843–1853
39. Epelbaum S, Paquet C, Hugon J et al (2019) How many patients are eligible for disease-modifying treatment in Alzheimer's disease? A French national observational study over 5 years. *BMJ Open* 9(6)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 8

Neuroimaging in Machine Learning for Brain Disorders

Ninon Burgos

Abstract

Medical imaging plays an important role in the detection, diagnosis, and treatment monitoring of brain disorders. Neuroimaging includes different modalities such as magnetic resonance imaging (MRI), X-ray computed tomography (CT), positron emission tomography (PET), or single-photon emission computed tomography (SPECT).

For each of these modalities, we will explain the basic principles of the technology, describe the type of information the images can provide, list the key processing steps necessary to extract features, and provide examples of their use in machine learning studies for brain disorders.

Key words Magnetic resonance imaging, Computed tomography, Positron emission tomography, Single-photon emission computed tomography, Neuroimaging, Medical imaging, Machine learning, Deep learning, Feature extraction, Preprocessing

1 Introduction

Medical imaging plays a key role in brain disorders. In clinical care, it is vital for detection, diagnosis, and treatment monitoring. It is also an essential tool for research to characterize the anatomical, functional, and molecular alterations in brain disorders, to better understand the pathophysiology, or to evaluate the effects of new treatments in clinical trials, for instance. Medical imaging of the brain is referred to as neuroimaging and involves different modalities such as X-ray computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), or single-photon emission computed tomography (SPECT).

Most neuroimaging modalities have been developed in the 1970s (Fig. 1). The first CT image of a brain was acquired in 1971 [1, 2]. This technology results from the discovery of X-rays by Wilhelm Röntgen in 1895 [3]. A few years later, PET [4] and then SPECT [5, 6] cameras were developed. Both modalities result from the discovery of natural radioactivity in 1896 by Henri Becquerel [7]. The first MR image of a brain goes back to 1978 [8]

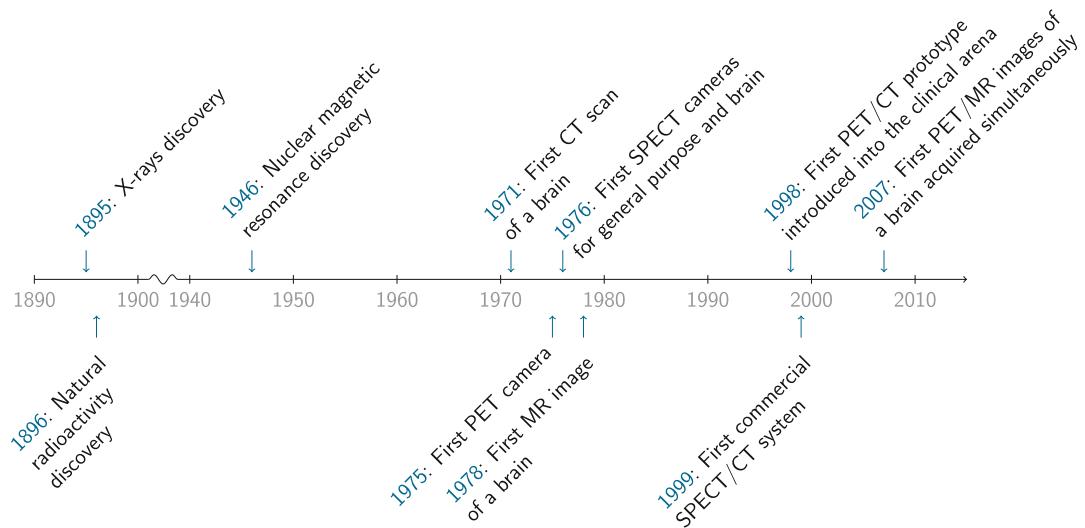


Fig. 1 Timeline of the main developments in neuroimaging

following the discovery of nuclear magnetic resonance in 1946 by Felix Bloch [9]. Some of these imaging modalities were later combined into hybrid scanners. The first prototype combining PET and CT was introduced into the clinical arena in 1998 [10], while the first PET and MR images of a brain simultaneously acquired were reported in 2007 [11, 12]. The first commercial SPECT/CT system dates back to 1999 [13], while SPECT/MR systems are still under development [14].

CT and MRI are the modalities of choice when studying brain anatomy, while SPECT and PET are used to image particular biological processes. Note that MRI is a versatile modality that allows studying both the structure and function of the brain, through the acquisition of different sequences. The use of these imaging modalities differs between clinical practice and research contexts. For example, CT is the main modality used in hospitals on adults [15], while MRI is by far the modality the most used for the study of brain disorders with machine learning (Fig. 2, top). The two most studied disorders with machine learning are brain tumors and dementia, mainly Alzheimer's disease (Fig. 2, bottom).

This chapter will start by shortly describing the nature of neuroimages, detailing the type of features that can be extracted from them, and listing software tools that can be used to do so. We will then briefly describe the principles of the imaging modalities the most used in machine learning studies: anatomical, diffusion, and functional MRI, CT, PET, and SPECT. For each modality, we will report the processing steps often performed to extract features, explain the type of information provided, and give examples of their use in machine learning studies.

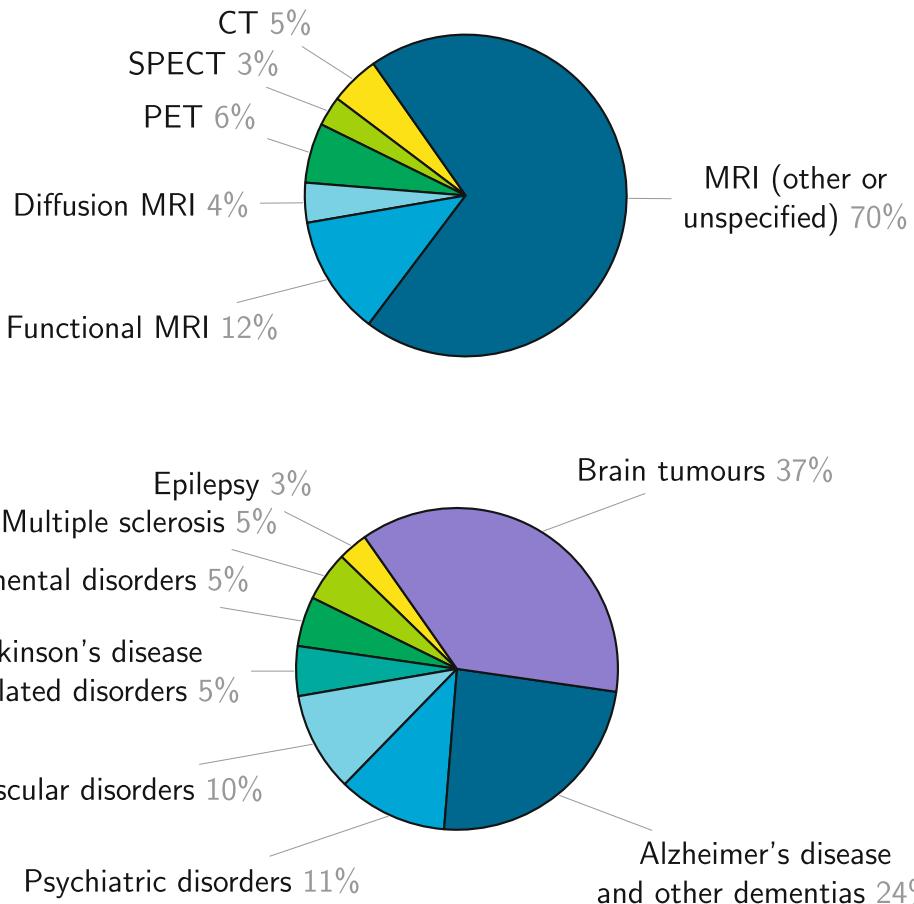


Fig. 2 Distribution by imaging modality (top) and brain disorder (bottom) of 1327 articles presenting a study using machine learning. Note that these numbers should only be taken as rough indicators as they result from a non-exhaustive literature search. The Scopus query and the resulting articles (after some manual filtering) are available as a public Zotero library (https://www.zotero.org/groups/4623150/neuroimaging_with_ml_for_brain_disorders/library)

2 Manipulating Neuroimages

In clinical routine, neuroimages are primarily exploited through visual inspection by a radiologist (or a neuroradiologist, who is a radiologist with an additional specialization in brain imaging, in expert hospitals) or a nuclear medicine physician. This results in a radiological report that is a written text describing the characteristics of the brain of the patient, its alterations, and possibly the most likely diagnosis. Note that neuroimaging exploration is usually requested by a neurologist or a psychiatrist and is associated with an indication that may correspond to the exploration of a set of symptoms (for instance, the exploration of a dementia syndrome) or to the confirmation of a potential diagnosis. Neuroimaging

alone will thus usually not provide a diagnosis. It will rather bring arguments in favor, or against, a potential diagnosis (for instance, in the exploration of a dementia syndrome, MRI can bring positive arguments for a diagnosis of Alzheimer's disease due to the observed atrophy in specific areas or on the contrary exclude this diagnosis by showing that the syndrome is due to a different cause such as a brain tumor). Overall, the diagnosis will generally be made by the neurologist or the psychiatrist based on a combination of clinical examination and a set of multimodal data (clinical and cognitive tests, radiological report, biomarkers, etc.).

However, the use of neuroimages goes way beyond visual inspection and is subject to quantification using image processing procedures. This is particularly true in research even though image processing tools are also increasingly used in clinical routine. A characteristic of these tools that differentiates them from general purpose image processing tools is their ability to handle three-dimensional (3D) images.

2.1 The Nature of 3D Medical Images

Most medical imaging devices acquire 3D images. This is the case of all the ones presented in this chapter (MRI, CT, PET, and SPECT). If 2D images are essentially 2D arrays of elements called pixels (for picture elements), 3D images are 3D arrays of elements called voxels (for volume elements). Depending on the imaging modality, voxel values will represent different properties of the underlying tissues. For example, in a CT image, they will be proportional to linear attenuation coefficients. The shape and size of a voxel will also depend on the imaging modality (or the type of sequence in MRI). When its three dimensions are of equal lengths, the voxel is isotropic; otherwise, it is anisotropic (*see* Fig. 3). For instance, a typical voxel size for a T1-weighted MR image is about $1 \times 1 \times 1 \text{ mm}^3$, while it is about $3 \times 3 \times 3 \text{ mm}^3$ for a functional MR image. Most neuroimaging modalities will have a voxel dimension between 0.5 mm and 5 mm.

Even though most neuroimages are 3D, they are visualized as 2D slices along different planes: axial, coronal, or sagittal (*see* Fig. 4). Multiple tools exist to visualize neuroimages. Several are available within suites such as FSLEYES,¹ Freeview,² or medInria,³ while others are independent such as Vinci,⁴ Mango,⁵ or Horos.⁶ Note that viewers may interpolate the images they display, which may be misleading (*see* Fig. 5 for an illustration).

¹ FSLEYES: <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLEYES>.

² Freeview: <https://surfer.nmr.mgh.harvard.edu/fswiki/FreeviewGuide>.

³ medInria: <https://med.inria.fr>.

⁴ Vinci: <https://vinci.sfmmpg.de>.

⁵ Mango: <http://ric.uthscsa.edu/mango>.

⁶ Horos: <https://horosproject.org>.

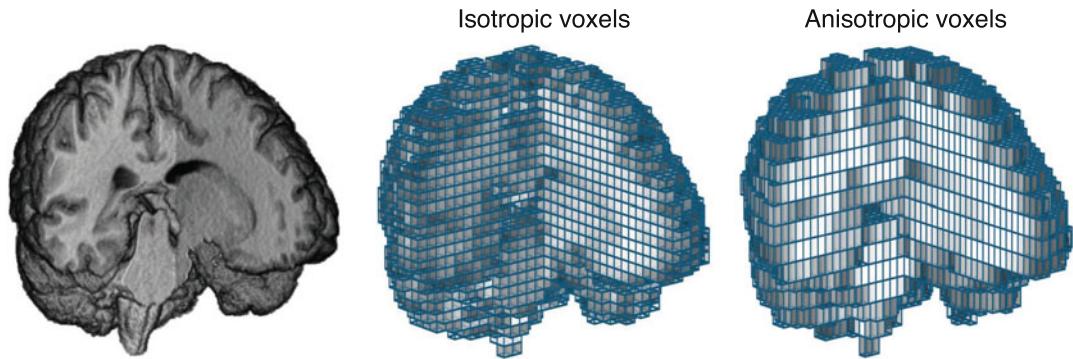


Fig. 3 Most neuroimaging modalities are three-dimensional. Left: volume rendering of an excavated T1-weighted MR image. Middle: voxel grid with isotropic, i.e., cubic, voxels overlaid on the MRI. Right: voxel grid with anisotropic, i.e., rectangular, voxels overlaid on the MRI

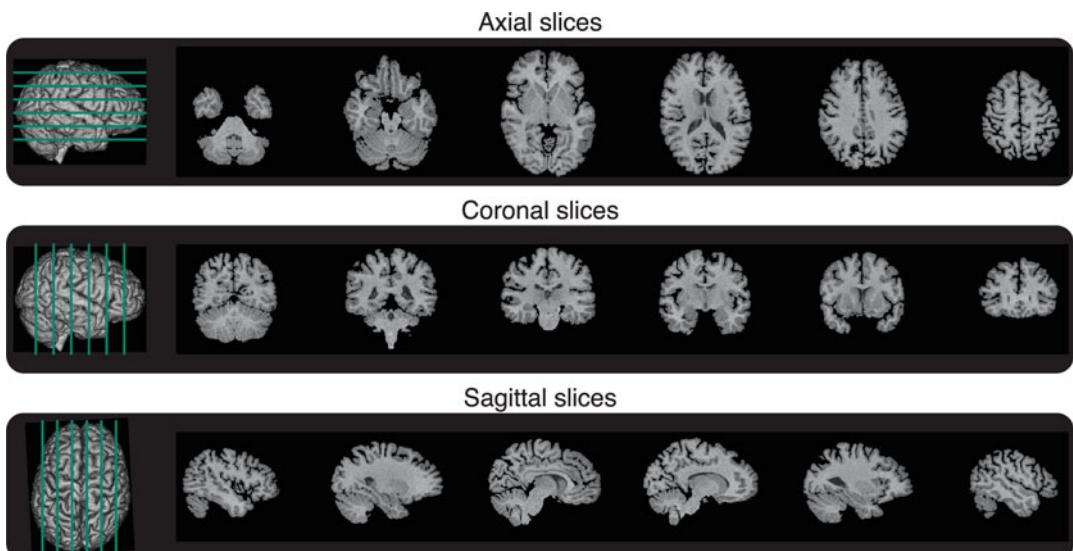


Fig. 4 Axial, coronal, and sagittal slices extracted from a T1-weighted MR image

2.2 Extracting Features from Neuroimages

When using machine learning to analyze images, one will often extract features. These features can be grouped into four categories that we will now describe and are illustrated in Fig. 6. Note that these features are conceptually the same for the different modalities but their actual content will differ (e.g., volume of a region for anatomical MRI vs average uptake in this region for PET). Modality-specific preprocessing and corrections often need to be applied before neuroimages can be analyzed; these will be described in Subheadings 3, 4, and 5.

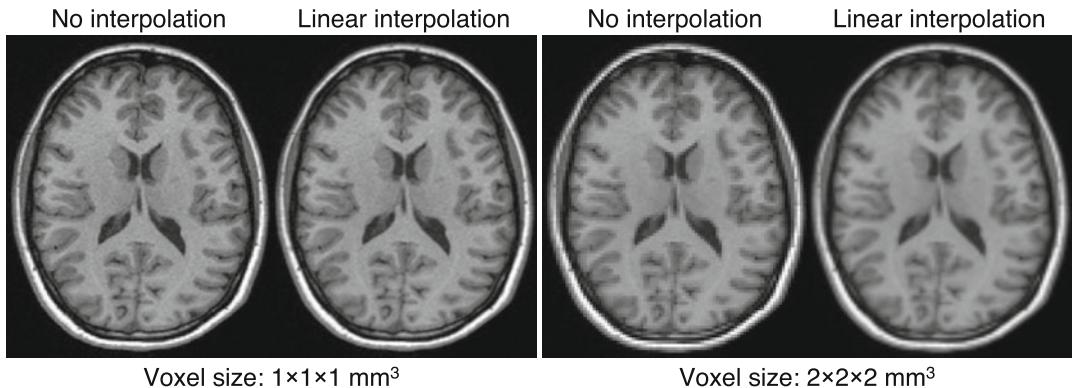
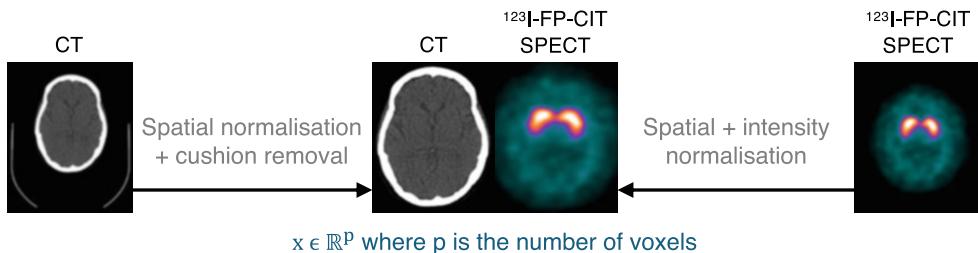


Fig. 5 Axial slice of a T1-weighted MRI with an isotropic voxel size originally of $1 \times 1 \times 1 \text{ mm}^3$ (left) and downsampled to $2 \times 2 \times 2 \text{ mm}^3$ (right) displayed without interpolation or with linear interpolation. If the difference with or without interpolation is subtle at $1 \times 1 \times 1 \text{ mm}^3$, it is well visible at $2 \times 2 \times 2 \text{ mm}^3$

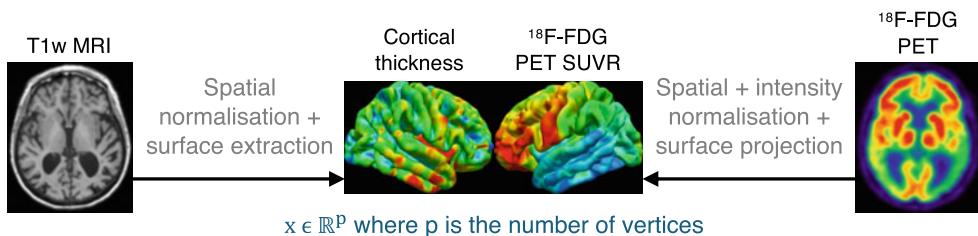
Voxel-Based Features As mentioned previously, all the imaging modalities described in this chapter produce volumetric images. The whole 3D image can be used as input of a machine learning algorithm. In that case, each subject is seen as a collection of values at each voxel of the image. These values can simply be the intensity of the image at each voxel after some minimal preprocessing (which is very often what is used in deep learning) or some more complex value extracted from the image (for instance, gray-level density from anatomical MRI; *see Subheading 3.1*). A prerequisite is often to align the images studied in a common space, by registering each image to a template and/or by performing a group-wise registration, thus guaranteeing a voxel-wise correspondence across subjects [16]. Note that this correspondence becomes particularly important when using a machine learning algorithm that takes as input a vector in which each element implicitly represents the same information for each subject (e.g., logistic regression or support vector machine).

Vertex-Based Features Studying the surface of the cortex is natural given its shape: it is a convoluted ribbon delimited by inner and outer surfaces. Moreover, surface-based characteristics can provide useful information such as for developmental or neurodegenerative diseases. Surfaces can be represented as meshes consisting of vertices, edges, and faces. The vertices encode position and properties such as cortical thickness. In the vertex-based feature scenario, each subject is seen as a collection of values at each vertex of the surface. Classical values computed at each vertex include cortical thickness and local surface area (*see Subheading 3.1*). As for voxel-based features, images studied are usually aligned in a common space to ensure a vertex-wise correspondence across subjects [17, 18].

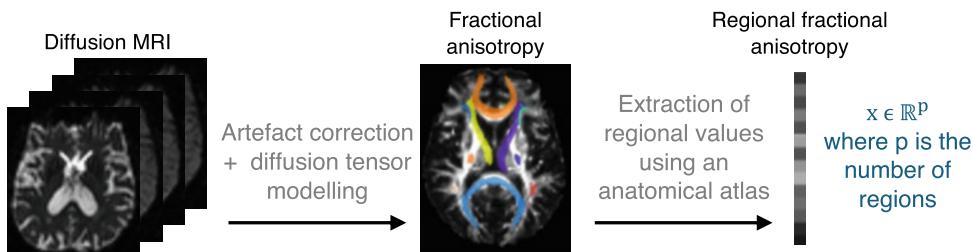
Voxel-based features



Vertex-based features



Regional features



Graph-based features

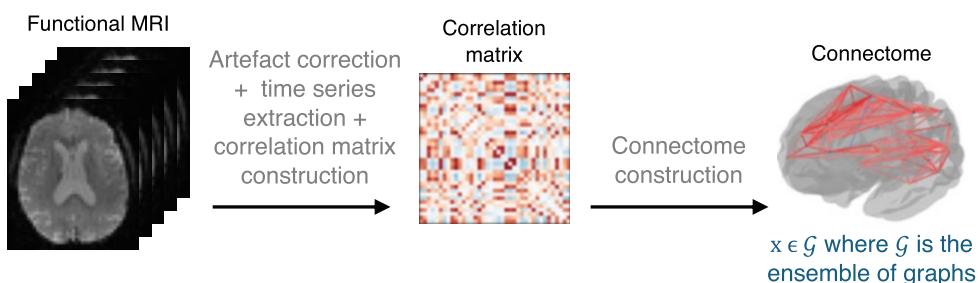


Fig. 6 Examples of voxel, vertex, regional, and graph features that can be extracted from neuroimages. It is, for instance, possible to extract voxel-based features from CT and SPECT images, vertex-based features from anatomical T1-weighted (T1w) MRI or PET images, regional features from diffusion MRI, and graph-based features from functional MRI. Note that the modalities are just examples. For instance, voxel-based features can be extracted for any modality. See Subheadings 3, 4, and 5 for a description of the imaging modalities

Regional Features The brain can be divided into subregions according to different criteria that can be anatomical or functional [16]. When considering regional features, each subject is seen as a collection of values for each region of the brain defined by an atlas. Many atlases exist, either anatomical or functional, with different degrees of granularity. A list can be found online.⁷ Classical values include the volume of a given region or the average image signal within a region.

Graph-Based Features A last way to represent an image is through a graph where nodes will correspond to brain regions and edges will encode a particular property (for instance, anatomical or functional connections, possibly together with their strength). Graphs can directly be used as features, but network indices characterizing global and local graph topology, e.g., efficiency or degree, can also be computed [19].

2.3 Neuroimaging Software Tools

The features described above can be obtained using neuroimaging software tools. However, an important step before any preprocessing or analysis is to properly organize data. The neuroimaging community proposed the Brain Imaging Data Structure [20], which specifies how to organize data in folders and sub-folders on disk and how to name the files. It also details the metadata necessary to describe neuroimaging experiments.

Many tools exist to process neuroimages.⁸ The historical generic frameworks include SPM⁹ [21], FSL¹⁰ [22], FreeSurfer¹¹ [23], or ANTs¹² [24]. Some tools are modality-specific such as MRtrix¹³ [25], dedicated to diffusion MRI, or AFNI¹⁴ [26], dedicated to functional MRI. Recent initiatives aim to make the use of neuroimaging tools easier by distributing them in containers (e.g., BIDSApps¹⁵ [27]), by providing in a single environment tools from preprocessing to machine learning (e.g., Nilearn¹⁶ [28]), or by providing automatic pipelines that do not require a particular

⁷ List of atlases: <https://www.lead-dbs.org/helpsupport/knowledge-base/atlassesresources>.

⁸ List of open source medical imaging software tools: <https://idoimaging.com>.

⁹ SPM: <https://www.fil.ion.ucl.ac.uk/spm>.

¹⁰ FSL: <https://fsl.fmrib.ox.ac.uk>.

¹¹ FreeSurfer: <https://surfer.nmr.mgh.harvard.edu>.

¹² ANTs: <http://stnava.github.io/ANTs>.

¹³ MRtrix: <https://www.mrtrix.org>.

¹⁴ AFNI: <https://afni.nimh.nih.gov>.

¹⁵ BIDSApps: <https://bids-apps.neuroimaging.io/apps>.

¹⁶ Nilearn: <https://nillearn.github.io>.

expertise in image processing (e.g., Clinica¹⁷ [29]). Other tools facilitate the application of deep learning approaches to neuroimages or medical images in general: for instance, MONAI,¹⁸ TorchIO¹⁹ [30], or ClinicaDL²⁰ [31].

3 Magnetic Resonance Imaging

Magnetic resonance imaging is the modality of choice to study brain anatomy, thanks to its high-resolution and excellent soft-tissue contrast, but the applications of MRI go well beyond studying anatomy. This technique can be used to examine tissue micro-architecture (diffusion MRI, covered in Subheading 3.2) or neuronal activity (functional MRI, covered in Subheading 3.3) but also to visualize the brain vasculature (MR angiography), study tissue perfusion and permeability (perfusion MRI), assess iron deposits and calcifications (susceptibility-based imaging), or measure the levels of different metabolites (MR spectroscopy). Note that MRI is an extremely versatile modality and that new sequences are constantly developed to study other brain characteristics.

3.1 Anatomical MRI

3.1.1 Basic Principles

In MRI, most images are obtained by exploiting a magnetic property, called spin, of the hydrogen atomic nuclei found in the water molecules present in the human body. In the absence of a strong external magnetic field, the directions of the proton's spins are random, thus cancelling each other out (Fig. 7a). When the spins enter a strong external magnetic field (B_0), they align either parallel or antiparallel, and they all precess around the B_0 axis, referred to as the z axis (Fig. 7b). As a result, they cancel each other out in the transverse (x, y) plane, but they add up along the z axis. The result of this vector addition, called net magnetization M_0 , is proportional to the proton density (Fig. 7c). With the application of a radio frequency pulse denoted as B_1 , the system of spins and the net magnetization are tipped by an angle determined by the strength and duration of the radio frequency pulse. For a 90° radio frequency pulse, the magnetization along the z axis (M_z) becomes zero and the magnetization in the transverse plane (M_{xy}) becomes equal to M_0 (Fig. 7d). As this radio frequency pulse provides energy, or excites, the spins, we also talk of radio frequency excitation.

¹⁷ Clinica: <https://www.clinica.run>.

¹⁸ MONAI: <https://monai.io>.

¹⁹ TorchIO: <https://torchio.readthedocs.io>.

²⁰ ClinicaDL: <https://clinicadl.readthedocs.io>.

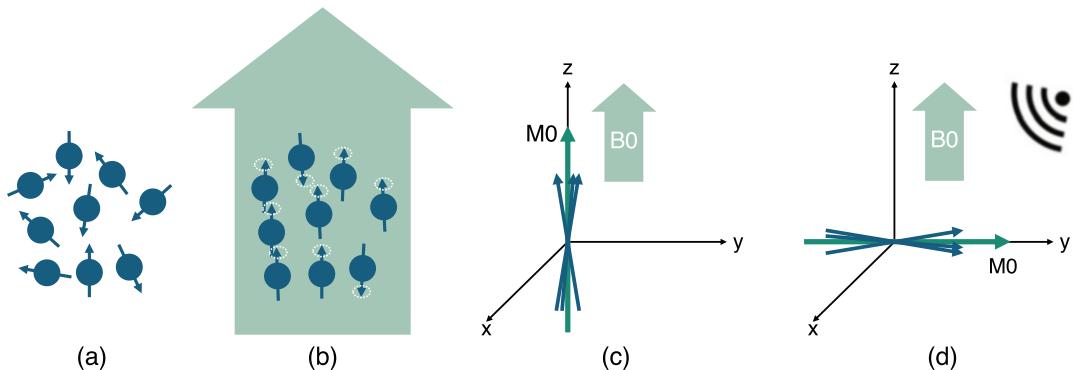


Fig. 7 MRI physics in a nutshell. (a) In the absence of a magnetic field, the directions of the proton's spins are random. (b) When the spins enter a strong external magnetic field (B_0), they align either parallel or antiparallel, and they all precess around the B_0 axis. (c) The net magnetization M_0 is proportional to the proton density. (d) With the application of a radio frequency pulse, the system of spins is tipped

When the radio frequency pulse is then turned off, two phenomena occur. First, the system of spins relaxes back to its preferred energy state of being parallel with B_0 in a time T_1 , called longitudinal or spin-lattice relaxation time, and the longitudinal magnetization M_z slowly recovers to its original magnitude M_0 . Second, each spin starts precessing at a frequency that is slightly different from the one of its neighboring spins because the field of the MRI scanner is not uniform and because each spin is influenced by the small magnetic fields of the neighboring spins. When the spins are completely dephased, they are evenly spread in the transverse plane, and M_{xy} becomes zero. M_{xy} decreases at a much faster rate than that at which M_z recovers to M_0 . The transverse relaxation time T_2 , also called spin-spin relaxation time, describes the M_{xy} decrease because of interference from neighboring spins, while T_{2^*} describes the decrease because of both spin-spin interactions and nonuniformities of B_0 . Finally, the MRI signal is obtained by measuring the transverse magnetization as an electrical current by induction.

The contrast in MR images depends on three main parameters: the proton density, the longitudinal relaxation time T_1 , and the transverse relaxation time T_2 . These parameters can be adjusted by changing the time at which the signal is recorded, called echo time, and the interval between successive excitation pulses, called repetition time. A T_1 -weighted image is created by choosing a short repetition time, a T_2 -weighted image by choosing a long echo time, and a proton density (PD)-weighted image by minimizing both T_1 and T_2 weighting of the image (long repetition time and short echo time). The corresponding images are referred to as T_1 -weighted MRI, T_2 -weighted MRI, and PD-weighted MRI. Note that many variations of these sequences exist (for instance, gradient-echo vs spin-echo) and the corresponding

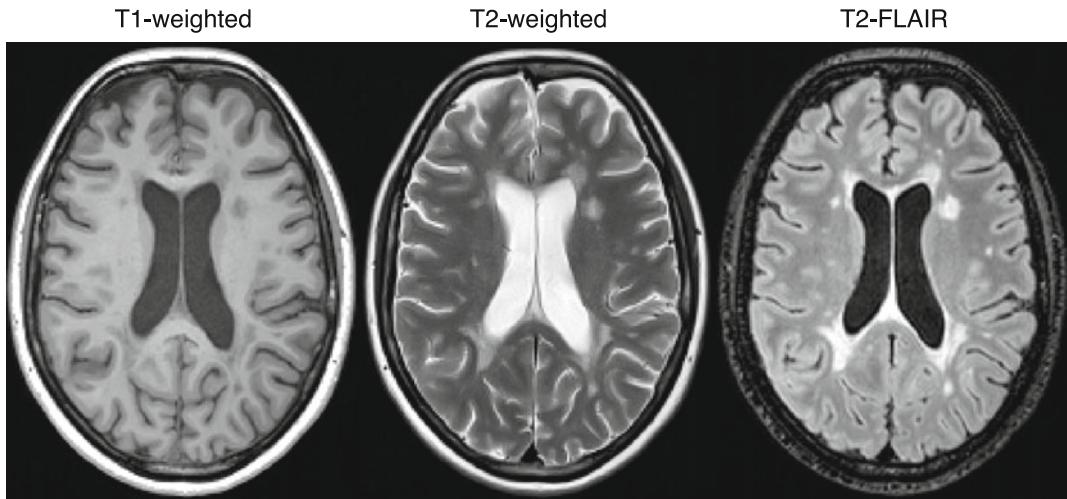


Fig. 8 Example of anatomical MR images. T1-weighted, T2-weighted, and T2-FLAIR images of a patient with multiple sclerosis from the MSSEG MICCAI 2016 challenge data set [32, 33]

implementation by different manufacturers usually comes with a specific commercial name (e.g., MPRAGE is a T1-weighted sequence available on Siemens scanners). Furthermore, many more anatomical sequences exist including T2*-weighted, T2-FLAIR (fluid-attenuated inversion recovery), or DIR (double inversion recovery). Examples are displayed in Fig. 8. The set of sequences chosen by the radiologist will depend on the potential disease that is being investigated. Some examples in the context of machine learning are given in Subheading 3.1.3.

3.1.2 Extracting Features from Anatomical MRI

Several preprocessing steps are often necessary before analyzing anatomical MR images to correct imperfections and ease their comparison.

Bias Field Correction MR images can be corrupted by a low-frequency and smooth signal caused by magnetic field inhomogeneity. This bias field induces variations in the intensity of the same tissue in different locations of the image, which deteriorates the performance of image analysis algorithms such as registration or segmentation. Several methods exist to correct these intensity inhomogeneities, the most popular being the N4 algorithm [34] available in ANTs [24].

Intensity Rescaling and Standardization As MRI is usually not a quantitative imaging modality, MR images can have different intensity ranges, and the intensity distribution of the same tissue type may be different between two images, which might affect the subsequent image preprocessing steps. The first point can be dealt with by globally rescaling the image, for example, between 0 and

1, using the minimum and maximum intensity values. More robust choices exist such as the z-score normalization (at each voxel, one subtracts the mean intensity of the image, and the result is divided by the standard deviation across the image), which can be made even more robust by only considering a percentile of the intensities for computing the mean and standard deviation. Intensity standardization, to solve the second point, can be achieved using techniques such as histogram matching [35].

Skull Stripping Extracranial tissues can be an obstacle for image analysis algorithms [36]. A large number of methods have been developed for brain extraction, also called skull stripping. Some are available in neuroimaging software platforms, such as FSL [22] or BrainSuite [37], and others as independent tools^{21,22} [38, 39]. Note that these methods can be sensitive to the presence of noise and artefacts, which can result in over or under segmentation of the brain.

Image Registration Medical image registration consists in spatially aligning two or more images, either globally (rigid and affine registration) or locally (nonrigid registration), so that voxels in corresponding positions contain comparable information. A large number of software tools have been developed for MRI-based registration [40]. They are available in all the major platforms (e.g., SPM [21], FSL [22], FreeSurfer [23], or ANTs [24]).

Image Segmentation Medical image segmentation consists in partitioning an image into a set of nonoverlapping regions. When processing brain images, these regions can correspond to tissue types, e.g., gray matter, white matter, and cerebrospinal fluid [41], but also to anatomical (e.g., hippocampus, pons) or functional (e.g., language network, sensorimotor network) regions defined by an atlas [42]. As for registration, many tools have been developed for MRI-based segmentation and are available, among others, in SPM [21], FSL [22], FreeSurfer [23], or ANTs [24].

Resulting Features Based on the combination of one, several, or all, of the previously mentioned preprocessing steps, various types of features can be extracted that correspond to those described in Subheading 2.2. For deep learning algorithms, which usually exploit voxel-based features, it is quite common to perform only very basic preprocessing. At the simplest, it can be intensity normalization (this step is mandatory for deep learning methods to work correctly). It is often combined with a bias field correction

²¹ HD-BET: <https://github.com/MIC-DKFZ/HD-BET>.

²² SynthStrip: <https://surfer.nmr.mgh.harvard.edu/docs/synthstrip>.

and a linear registration to a common space. Another common type of voxel-based features is that of tissue density maps (e.g., gray matter or white matter density) [43]. Their extraction involves bias field correction, registration to a common space, and tissue segmentation. Common vertex-based features are the local thickness and the local surface area [44]. Regional features are usually the volume of different regions of the brain, but they can also be the average intensity within the region or the average of another image-derived value. They can as well be related to lesions (for instance, the volume of multiple sclerosis lesions or of different compartments of a brain tumor) rather than anatomical regions. Finally, graph-based features can also be computed from anatomical MRI [45] even though this representation is more common for diffusion MRI and functional MRI.

3.1.3 Examples of Applications in Machine Learning Studies

T1-weighted MRI is the sequence the most commonly found in machine learning studies applied to brain disorders. Several features can be extracted from T1-weighted MRI such as the volume of the whole brain or of regions of interest; the density of a particular tissue, e.g., gray matter; or the local cortical thickness and surface area. All these features, as well as the raw T1-weighted MR images, have, for example, largely been used for the computer-aided diagnosis of dementia, in particular Alzheimer's disease, as they highlight atrophy, i.e., the neuronal loss that is a marker of neurodegenerative diseases [46–49].

T1-weighted MR images acquired with and without the injection of a contrast agent are often used in the context of brain tumor detection and segmentation, progression assessment, and survival prediction as they allow distinguishing active tumor structures [50]. Such tasks also typically rely on another sequence called T2-weighted fluid-attenuated inversion recovery (T2-FLAIR) that allows visualizing a wide range of lesions on top of tumors [51], such as those appearing with multiple sclerosis [52, 53] or age-related white matter hyperintensities (also called leukoaraiosis, which is linked to small vessel disease).

3.2 Diffusion-Weighted MRI

3.2.1 Basic Principles

Diffusion MRI [54, 55] allows visualizing tissue micro-architecture, thanks to the diffusion of water molecules. Depending on their surroundings, water molecules are able to either move freely, e.g., in the extracellular space, or move following surrounding constraints, e.g., within a neuron. In the former situation, the diffusion is isotropic, while in the later it is anisotropic. Contrast in a diffusion MR image originates from the fact that following the application of an excitation pulse, water molecules that move in a particular direction, and so the protons they contain, do not have the same magnetic properties as the ones that move randomly but not far from their origin point. The excitation pulse is parametrized by a weighting coefficient b : the higher the b -value, the more

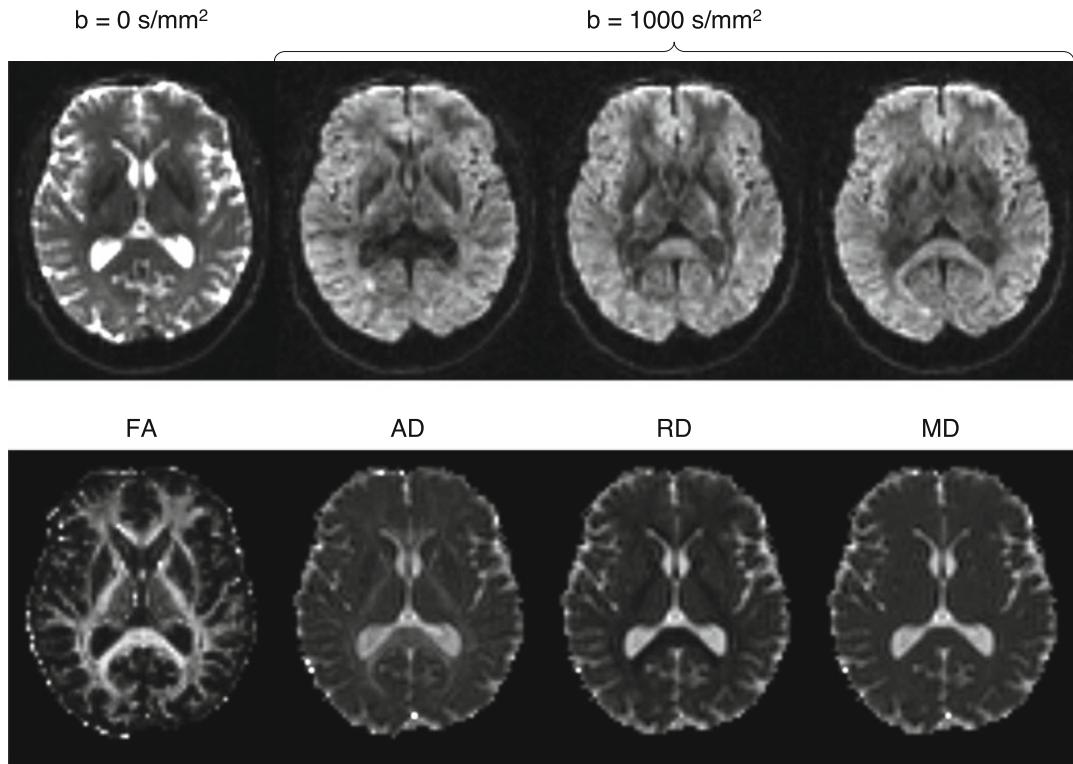


Fig. 9 Example of diffusion-weighted MR images. Top: diffusion volumes acquired using different b -values (0 and 1000 s/mm^2) and gradient directions. Bottom: parametric maps resulting from diffusion tensor modeling (fractional anisotropy, FA; axial diffusivity, AD; radial diffusivity, RD; and mean diffusivity, MD)

sensitive the acquisition is to water diffusion, but the lower the signal-to-noise ratio. Several diffusion MRI volumes, each volume corresponding to a particular b -value and gradient direction, are usually acquired. See examples in Fig. 9 (top row).

3.2.2 Extracting Features from Diffusion MRI

Diffusion MR images are typically acquired with echo-planar imaging, a technique that spatially encodes the MRI signal in a way that enables fast acquisitions with a relatively high signal-to-noise ratio. However, echo-planar imaging induces geometric distortions and signal losses known as magnetic susceptibility artifacts. Other artifacts include eddy currents (due to the rapid switching of diffusion gradients), intensity inhomogeneities (as for anatomical MRI), and potential movements of the subject during the acquisition. These artifacts need to be corrected before further analyzing the images. Various methods exist to do so; they are reviewed in [56]. Two widely used tools enabling the preprocessing of diffusion MR images are FSL [22] and MRtrix [25], but others exist²³ [56].

²³ List of tools and software packages to process diffusion MRI: <https://github.com/dmriprep/processing/neuroimage-review-2022>.

Once artifacts have been corrected, diffusion MR images can be analyzed in different ways. One of the earliest strategy for modeling water diffusion is the diffusion tensor imaging (DTI) model [57]. Such model can output parametric maps describing several diffusion properties: fractional anisotropy (FA, directional preference of diffusion), mean diffusivity (MD, overall diffusion rate, also called apparent diffusion coefficient), axial diffusivity (AD, diffusion rate along the main axis of diffusion), and radial diffusivity (RD, diffusion rate in the transverse direction). Examples of parametric maps are displayed in Fig. 9 (bottom row). DTI tractography [58] goes one step further by reconstructing white matter tracts. Other diffusion models have been developed to better characterize tissue micro-architecture. This is, for example, the case of neurite orientation dispersion and density imaging (NODDI) [59], which enables the study of neurite morphology by disentangling neurite density and orientation dispersion that both independently influence fractional anisotropy.

One can then again compute most of the different types of features covered in Subheading 2.2. Voxel-based features will represent the value of a given parametric map (e.g., FA, MD). Surface-based features are seldom used because diffusion MRI often focuses on the white matter even though it is in principle possible to project maps that are of interest in the gray matter onto the cortical surface. Regional features represent the average of a given map (e.g., FA, MD) in a set of anatomical regions. Graph-based features can be computed as follows, vertices are often regions of the cortex, and edges correspond to the connection strength, which can be derived, for instance, from the number of tracts connecting two regions or the average FA within those tracts.

3.2.3 Examples of Applications in Machine Learning Studies

Machine learning studies have mainly used diffusion MRI to assess white matter integrity. This has been done in a very wide variety of disorders. For example, fractional anisotropy and mean diffusivity have been used to differentiate cognitively normal subjects from patients with mild cognitive impairment or Alzheimer's disease [60, 61]. Diffusion MRI has also been exploited to perform tumor grading or subtyping [62] following the assumption that the cellular structure may differ between cancerous and healthy tissues.

3.3 Functional MRI

3.3.1 Basic Principles

When a region of the brain gets activated by a cognitive task, two phenomena occur: a local increase in cerebral blood flow and changes in oxygenation concentration [63]. Functional MRI (fMRI) is used to measure the latter phenomenon. The blood-oxygen-level-dependent (BOLD) contrast originates from the fact that hemoglobin molecules that carry oxygen have different magnetic properties than hemoglobin molecules that do not carry oxygen.

Task fMRI consists in inducing particular neural states, for example, by performing tasks involving the visual or auditory systems and then comparing the signals recorded during the different states. As the differences observed are small, it is important to preserve at best the signal-to-noise ratio that could be degraded because of head motion or polluted by fluctuations of the cardiac and respiratory cycles. This is done by quickly acquiring multiple image volumes with echo-planar imaging. The BOLD signal also varies when the brain is not performing any particular task [64]. These spontaneous fluctuations are studied with resting-state fMRI.

3.3.2 Extracting Features from Functional MRI

The preprocessing of functional MR images has two main objectives: limit the effect of nonneural sources of variability and correct acquisition-related artifacts [65]. Preprocessing steps can, for example, include susceptibility distortion correction (as for diffusion MRI); head motion correction, by registering each volume in the time series to a reference volume (e.g., the first volume); slice-timing correction, to eliminate differences between the time of acquisition of each slice in the volume; or physiologic noise correction, by temporal filtering [63, 65]. These preprocessing steps can be performed using tools such as SPM [21], FSL [22], or AFNI [26], but also using the dedicated fMRIPrep workflow [65].

The majority of machine learning studies in brain disorders focuses on resting-state rather than task fMRI [66]. This can be explained by the fact that the resting-state protocol is simpler and allows multi-site studies (as it is less sensitive to changes in local experimental settings) [66], which should result in larger samples. Depending on the application, preprocessed resting-state fMRI data may be further processed to extract features. One can directly use voxel-based features (or vertex-based features by projecting the functional MRI signal onto the cortical surface) [67]. Nevertheless, to the best of our knowledge, the most common features are graph-based. Indeed, most supervised algorithms for classification or regression use brain networks extracted from resting-state time series. In these networks, also called connectomes, the vertices correspond to brain regions, which size may vary, and the edges encode the functional connectivity strength, which corresponds to the correlation between time series.

3.3.3 Examples of Applications in Machine Learning Studies

Machine learning methods exploiting resting-state fMRI data have been used to investigate brain development and aging, but also neurodegenerative and psychiatric disorders [66]. Functional connectivity patterns have, for instance, been used to distinguish patients with schizophrenia from healthy controls [68] or discriminate schizophrenia and bipolar disorder from healthy controls [69].

4 X-Ray Imaging

X-ray imaging is built on the work of Röntgen who observed that if a “hand be held before the fluorescent screen, the shadow shows the bones darkly, with only faint outlines of the surrounding tissues” [3].

4.1 X-Ray and Angiography

When an X-ray beam passes through the body, part of its energy is absorbed or scattered: the number of X-ray photons is reduced by attenuation (Fig. 10, left). On the opposite side of the body, detectors capture the remaining X-ray photons, and an image is generated. In an X-ray image, the contrast, defined as the relative intensity change produced by an object, originates from the variations in linear attenuation coefficient with tissue type and density.

X-ray imaging provides excellent contrast between bone, air, and soft tissue but very little contrast between the different types of soft tissue, hence its limited use when studying brain disorders. However, coupled with the injection of an iodine-based contrast agent, X-ray imaging enables visualizing cerebral blood vessels and detecting potential abnormalities such as an aneurysm. This technique is called X-ray angiography.

4.2 Computed Tomography

4.2.1 Basic Principles

Although the X-ray images produced were originally in 2D, X-ray computed tomography enables the reconstruction of 3D images by rotating the X-ray source and detectors around the body (Fig. 10, right). Rather than using the absolute values of the linear attenuation coefficients, CT image intensities are expressed in a standard

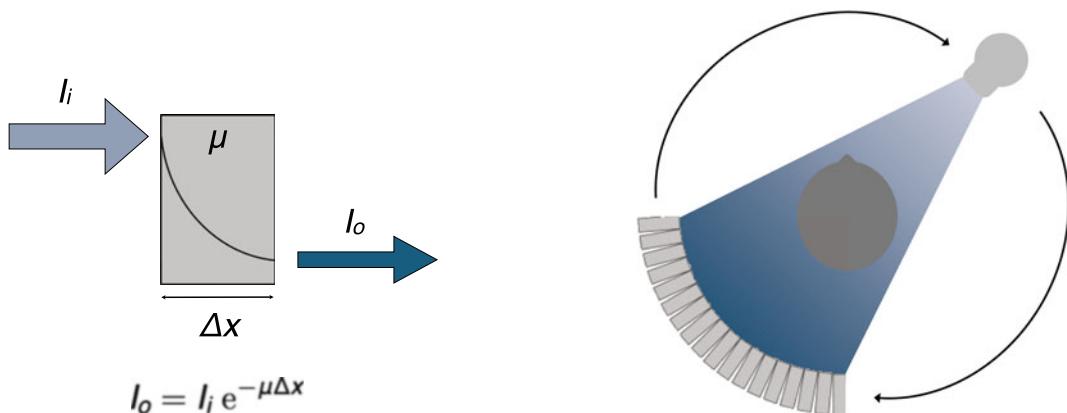


Fig. 10 Left: attenuation of X-rays by matter. As it passes through a material of thickness Δx and linear attenuation coefficient μ , the X-ray beam is attenuated. Its intensity decreases exponentially with the distance travelled: $I_o = I_i e^{-\mu \Delta x}$, where I_i and I_o are the input and output X-ray intensities. Right: third-generation CT. A 3D image is created by rotating the X-ray source and detectors around the body

unit, the Hounsfield unit (HU). The tissue attenuation coefficient is compared to the attenuation value of water and displayed on the Hounsfield scale:

$$x_{HU} = 1000 \times \frac{\mu_{\text{tissue}} - \mu_{\text{water}}}{\mu_{\text{water}} - \mu_{\text{air}}}$$

where μ_{water} and μ_{air} are the linear attenuation coefficients of water and air, respectively. For example, air has an attenuation of -1000 HU, water of 0 HU, and cortical bone between 500 and 1900 HU.

As for 2D X-ray imaging, the injection of an iodine-based contrast agent improves the visualization of cerebral blood vessels. This technique, called CT angiography, is not the only one relying on a contrast agent. CT perfusion tracks the bolus of contrast agent over time and measures the resulting change in signal intensity. Perfusion parameters such as the cerebral blood flow or volume can then be derived [70].

4.2.2 Extracting Features from CT Images

Contrary to MRI, CT images usually do not require extensive pre-processing steps [71]. It can however be useful to extract the head from the hardware elements visible on the image (e.g., the bed or pillow) or extract the brain. This can be done using thresholding and morphological operators. Another common step is spatial normalization.

In the context of stroke, non-contrast CT is useful to detect an intracranial hemorrhage, which appears brighter than the surrounding tissues, or to estimate the extent of early ischemic injury, which results in a loss of gray-white matter differentiation. CT angiography can help identify a potential intracranial arterial occlusion, and CT perfusion allows differentiating the regions with nonviable/non-salvageable tissue, which have very low cerebral blood flow and volume, from the viable and potentially salvageable regions [70]. These techniques may also be employed in the context of brain tumors. In particular, contrast-enhanced CT can detect areas presenting a blood-brain barrier breakdown [72]. An example of CT acquired before and after contrast injection is displayed in Fig. 11.

To the best of our knowledge, CT is most often used in machine learning in the form of voxel-based features (the image intensities after some minimal preprocessing steps).

4.2.3 Examples of Applications in Machine Learning Studies

The vast majority of machine learning studies relying on CT images, particularly non-contrast CT, focus on cerebrovascular disorders [73, 74]. Non-contrast CT images were, for example, used for the detection of intracranial hemorrhage and its five subtypes [75]. A first neural network was in charge of identifying the presence or absence of intracranial hemorrhage and a second of determining the intracranial hemorrhage subtype, which depends

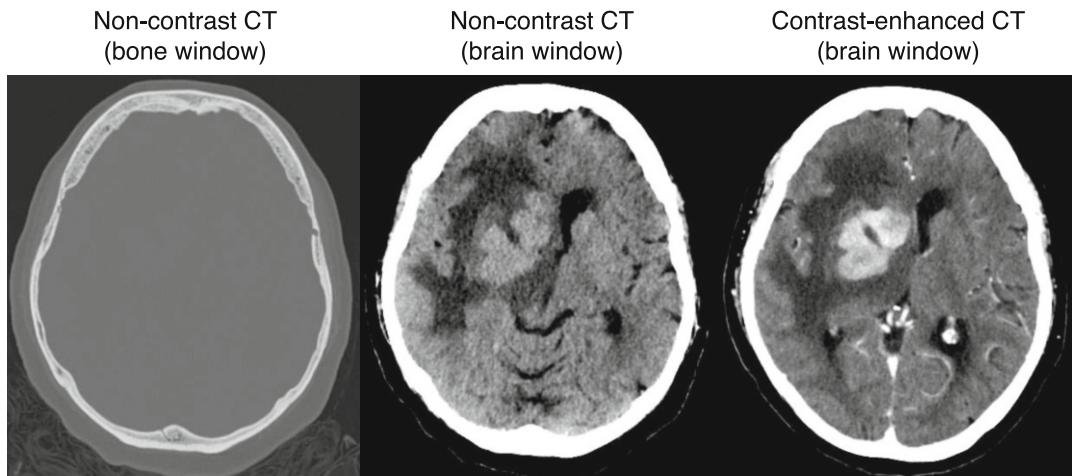


Fig. 11 Example of CT images. Non-contrast CT images, whose window levels were adjusted to better visualize bone or brain tissues and contrast-enhanced CT image of a patient with lymphoma. Case courtesy of Dr Yair Glick, [Radiopaedia.org](https://www.radiopaedia.org/cases/94844), rID: 94844

on the bleeding location [75]. In [76], non-contrast CT and CT perfusion images were used to segment the core of stroke lesions, as the lesion volume is a key measurement to assess the prognosis of acute ischemic stroke patients.

5 Nuclear Imaging

In X-ray CT imaging, the photons that are detected originate from an X-ray source. In nuclear imaging, and more precisely emission computed tomography, the photons detected are emitted from a radiopharmaceutical that has been intravenously injected to the patient.

5.1 Positron Emission Tomography

5.1.1 Basic Principles

Positron emission tomography is an imaging technique that requires the injection of a substance labeled with a positron-emitting radioactive isotope [77]. The labeled substance is distributed throughout the patient's body by the blood circulation and accumulates in target regions. The positrons emitted by the radioactive isotope combine with the electrons present in the tissues and annihilate. Each annihilation produces two nearly collinear photons (Fig. 12). The two photons are simultaneously detected by two opposing detectors, and a coincidence event is assigned to a line of response connecting the two detectors.

Note that the most common isotope in clinical routine is fluorine-18 (^{18}F), which has the advantage of a relatively long half-life (110 min) and thus does not require the presence of a cyclotron at the scanning site. Nevertheless, other isotopes are

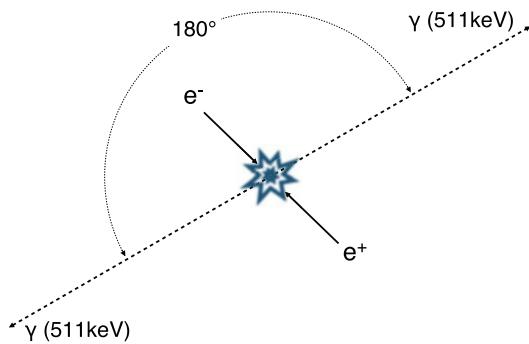


Fig. 12 PET annihilation. When a positron (e^+) and an electron (e^-) collide, they annihilate and create a pair of collinear gamma rays (γ)

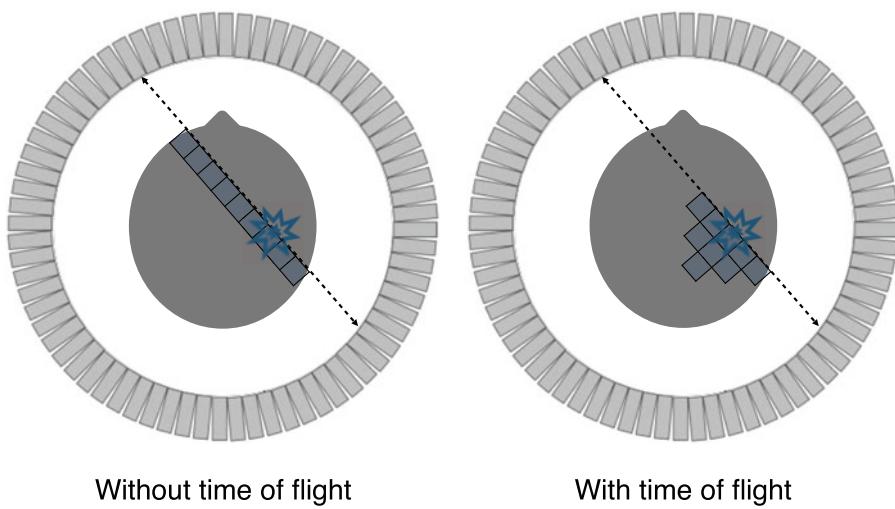


Fig. 13 Illustration of PET data detection. Without time-of-flight, the annihilation is located with equal probability along the line of response, while with time-of-flight it is located in a limited portion of the line of response

used. In particular, carbon-11 (^{11}C), which has a shorter half-life (20 min), is often used in research facilities equipped with a cyclotron.

In a time-of-flight PET system, the difference in arrival times between the two coincident photons is measured. Without time-of-flight information, the annihilation is located with equal probability along the line of response, while with time-of-flight information, the annihilation site can be reduced to a limited range (Fig. 13), thus decreasing the spatial uncertainty and increasing the signal-to-noise ratio. Once reconstructed, the PET image is a map of the radioactivity distribution throughout the body.

Two main protocols exist when acquiring PET data. Most acquisitions are static: the radiotracer is injected several minutes before the acquisition (e.g., between 30 and 60 min), which gives

the tracer time to diffuse in the body and accumulate in the target regions. The subject is then placed in the scanner and the acquisition lasts typically around 15 min. In the dynamic protocol, the subject is first installed in the scanner, and the acquisition starts at the same time the tracer is being injected. This allows recording how the tracer diffuses in the body. Dynamic acquisitions are less common than static ones because of their duration of 60–90 min, which reduces patient throughput. In both static and dynamic protocols, the acquisition is often split in frames of fix (in the static case) or increasing (in the dynamic case) duration. A static acquisition of 15 min can typically be split into three frames of 5 min, resulting in three PET volumes, each corresponding to the average amount of radioactivity detected at each voxel during the time frame.

^{18}F -fluorodeoxyglucose (FDG) is the most widely used PET radiopharmaceutical [77, 78]. As an analogue of glucose, FDG is transported to a cell, but, unlike glucose, it remains trapped in the cell. This radiopharmaceutical is an excellent marker of changes in glucose metabolism. In the brain, FDG acts as an indirect marker of synaptic dysfunction and is part of the diagnosis of epilepsy and neurodegenerative diseases, such as Alzheimer's disease [79].

If ^{18}F -FDG is a nonspecific tracer, other radiopharmaceuticals target specific molecular or biological processes and are thus preferentially used for studying specific diseases. Amyloid tracers, such as the ^{11}C Pittsburgh compound B, ^{18}F -florbetapir, ^{18}F -florbetaben and ^{18}F -flutemetamol, which bind to fibrillar $\text{A}\beta$ plaques, or tau tracers, such as ^{18}F -flortaucipir, and ^{18}F -MK-6240, which bind to neurofibrillary tangles, are, for example, used in the diagnosis of dementia syndromes [80]. Examples are displayed in Fig. 14. Of note, the so-called amyloid tracers are in fact not specific of amyloid and also bind to myelin in the white matter, making them of

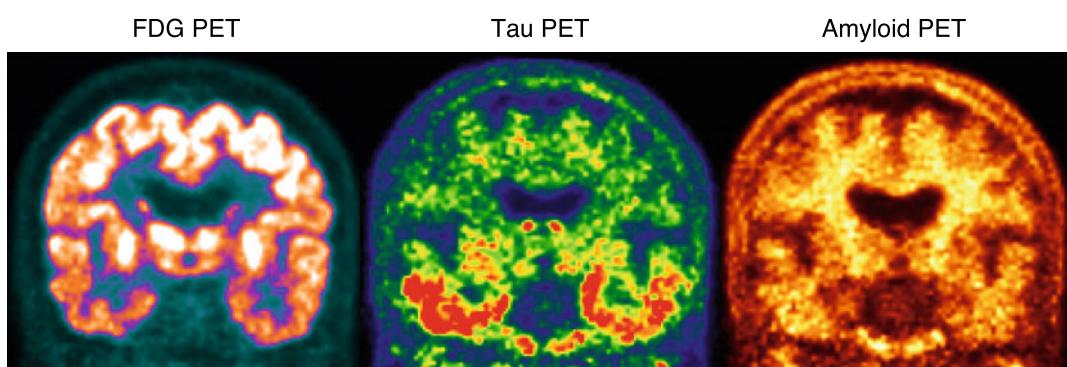


Fig. 14 Example of PET images. Left: ^{18}F -FDG PET displaying brain glucose metabolism. Middle: ^{18}F -flortaucipir PET displaying the presence of tau neurofibrillary tangles. Right: ^{18}F -florbetapir PET displaying the presence of amyloid plaques. All the images correspond to the same Alzheimer's disease patient from the ADNI study [83]

interest for demyelinating disorders such as multiple sclerosis [81]. ^{11}C -methionine and ^{18}F -fluoroethyltyrosine are both used in neuro-oncology [82]. Note that these are just examples of tracers and dozens of tracers exist for imaging specific molecular or biological processes.

5.1.2 Extracting Features from PET Images

The reconstruction procedure of the PET signal already includes several corrections (e.g., attenuation and scatter corrections), but several processing steps can be performed before further analyzing PET images. The first one is often motion correction. This is typically done by rigidly registering each frame to a reference frame. The registered frames are then averaged to form a single volume. To allow for intersubject comparison, brain PET images need to be intensity normalized, for example, to compensate for variations in the patients' weight or dose injected. Standardized uptake value ratios (SUVRs) are generated by dividing a PET image by the mean uptake in a reference region. This region can be obtained from an atlas, and in this case chosen depending on the tracer and disorder suspected, or in a data-driven manner [84]. Partial volume correction can be performed to limit the spill out of activity outside of the region where the tracer is meant to accumulate [85] using tools such as PETPVC [86]. Finally, PET images can also be spatially normalized. If an anatomical image (preferably MRI but also CT) of the subject is available, the PET image is rigidly registered to the anatomical image, and the anatomical image is registered to a template, often in standard space. By composing the two transformations, the PET image is spatially normalized. Alternatively, if no anatomical image is available, the PET image can directly be registered to a PET template, for example, as implemented in SPM [87]. Dynamic PET images are further processed to extract quantitative physiological data using kinetic modeling, which is introduced in [77, 78].

One can then obtain different types of features, as described in Subheading 2.2. Voxel-based features will very often be the SUVR at each voxel, usually after spatial normalization. Vertex-based features will generally be the SUVR projected onto the cortical surface [88]. Regional features will usually correspond to the average SUVR in each region of a parcellation. Graph-based features are less used than for diffusion or functional MRI but are still employed to study the so-called metabolic connectivity [89].

5.1.3 Examples of Applications in Machine Learning Studies

Machine learning studies have mainly exploited brain PET images in the context of dementia [90]. For example, the usefulness of ^{18}F -FDG PET to differentiate patients with Alzheimer's disease from healthy controls and patients with stable mild cognitive impairment from those who subsequently progressed to Alzheimer's disease has been shown in [48, 91, 92]. ^{18}F -FDG PET has also been used to differentiate frontotemporal dementia from

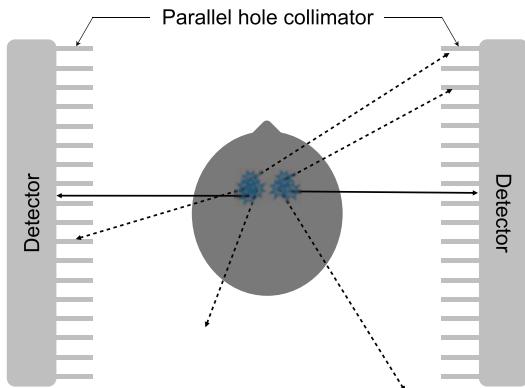


Fig. 15 Illustration of a two-head SPECT system with a parallel hole collimator. The photons whose emission direction is perpendicular to the detector heads have a higher probability of being detected (solid lines)

Alzheimer's disease [93]. In neuro-oncology, ^{11}C -methionine has been used to predict glioma survival [94] or to differentiate recurrent brain tumor from radiation necrosis [95].

5.2 Single-Photon Emission Computed Tomography

5.2.1 Basic Principles

Single-photon emission computed tomography is an imaging technique that requires the injection of a substance labeled with an isotope that directly emits gamma radiation. Typical isotopes employed in neurology are technetium-99m ($^{99\text{m}}\text{Tc}$) and iodine-123 (^{123}I). As for PET, the labeled substance is distributed throughout the patient's body by the blood circulation and accumulates in target regions. The photons emitted are detected by one to three detector heads, called gamma cameras, that rotate around the patient. Having multiple heads allows reducing image acquisition time and improving sensitivity as more photons can be detected. Collimators are placed in front of the detector heads to localize the origin of the gamma rays: a gamma ray moving from the patient toward the camera has a higher probability of being detected if its direction aligns with the collimator (Fig. 15) [96]. Once reconstructed, the SPECT image is a map of the radioactivity distribution throughout the body. Both dynamic and static protocols exist when acquiring SPECT data.

SPECT is able to visualize and quantify changes in cerebral blood flow and neurotransmitter systems, such as the dopamine system [97, 98]. To image cerebral blood flow, the two most widely used tracers are $^{99\text{m}}\text{Tc}$ -HMPAO and $^{99\text{m}}\text{Tc}$ -ECD [97, 99]. These tracers can, for example, be employed in the context of dementia as a decrease in neural function will result in a decrease in cerebral blood flow in different regions. SPECT plays a key role when studying Parkinsonian syndromes, which are characterized by a loss of dopaminergic neurons. In this context, tracers targeting the dopaminergic system, such as ^{123}I - β -CIT and ^{123}I -FP-CIT

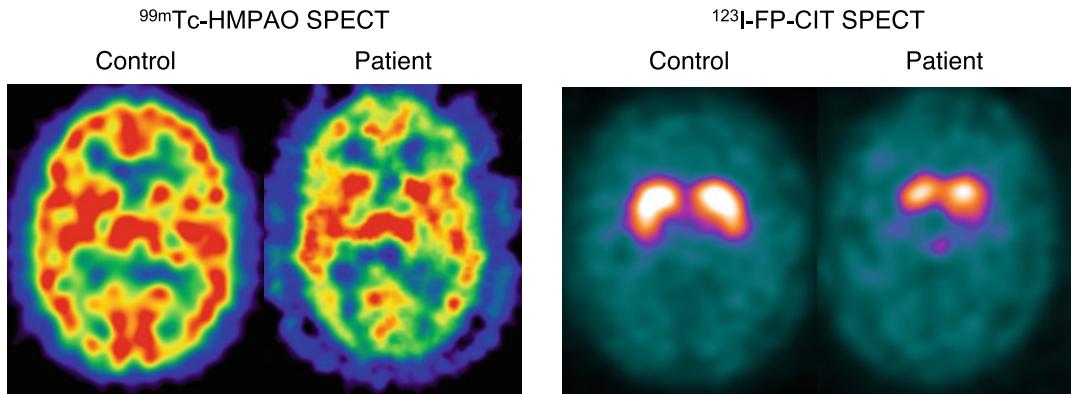


Fig. 16 Examples of SPECT images. Left: $^{99\text{m}}\text{Tc}$ -HMPAO SPECT images of a normal control and an epileptic patient (<http://spect.yale.edu>) [100]. Right: ^{123}I -FP-CIT SPECT images of a normal control and a patient with Parkinson's disease from the PPMI study [101]

(also called DaTscan), are employed to differentiate essential tremor from neurodegenerative Parkinsonian syndromes or distinguish dementia with Lewy bodies from other dementias [98]. Examples of SPECT images are displayed in Fig. 16.

5.2.2 Extracting Features from SPECT Images

After the reconstruction of a SPECT image, which includes several corrections, two processing steps are typically performed: intensity normalization and spatial normalization [97, 98]. As for PET, the intensity of a SPECT image can be normalized using a reference region, and the image can be spatially normalized by directly registering it to a SPECT template or by registering it first to an anatomical image.

As for PET, the most common feature types are voxel-based (the normalized signal at each voxel) and regional features (often the average normalized signal within a region). To the best of our knowledge, vertex-based and graph-based features are rarely used although they could in principle be computed.

5.2.3 Examples of Applications in Machine Learning Studies

Machine learning studies have mainly exploited brain SPECT images for the computer-aided diagnosis of Parkinsonian syndromes [102]. ^{123}I -FP-CIT SPECT was, for instance, used to distinguish Parkinson's disease from healthy controls [103, 104], predict future motor severity [105], discriminate Parkinson's disease from non-Parkinsonian tremor [104], or identify patients clinically diagnosed with Parkinson's disease but who have scans without evidence of dopaminergic deficit [104].

In studies targeting dementia, both $^{99\text{m}}\text{Tc}$ -HMPAO [106] and $^{99\text{m}}\text{Tc}$ -ECD [107] tracers were used to differentiate between images from healthy subjects and images from Alzheimer's disease patients.

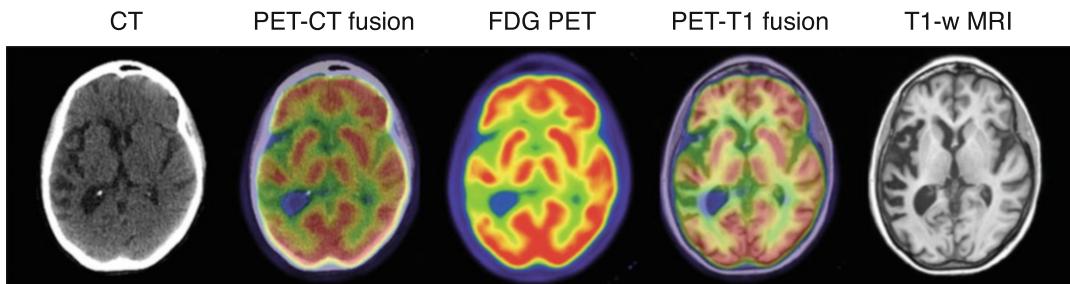


Fig. 17 Example of ^{18}F -FDG PET, CT, T1-weighted MRI, and fused images

6 Conclusion

Neuroimaging plays a key role for the study of brain disorders. If some modalities provide information regarding the anatomy of the brain (CT and MRI), others provide functional or molecular information (MRI, PET, and SPECT). To provide a complete picture of biological processes and their alterations, it is often necessary to combine multiple brain imaging modalities (Fig. 17). This can be done by acquiring images with multiple standalone systems or with hybrid systems such as SPECT/CT, PET/CT, or PET/MRI scanners [108].

When analyzing neuroimages, both modality-specific and modality-agnostic processing steps must often be performed. These should be performed with care to obtain reliable features. Machine learning and deep learning are widely used to analyze neuroimaging data. The most common tasks are classification for computer-aided diagnosis, prognosis and disease subtyping, and segmentation to characterize anatomical structures and lesions.

Acknowledgements

The author would like to thank the editor for useful suggestions. The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAHU-0006 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6).

References

1. Ambrose J (1973) Computerized transverse axial scanning (tomography): Part 2. Clinical application. *Br J Radiol* 46(552):1023–1047. <https://doi.org/10.1259/0007-1285-46-552-1023>
2. Hounsfield GN (1973) Computerized transverse axial scanning (tomography): Part 1. Description of system. *Br J Radiol* 46(552):1016–1022. <https://doi.org/10.1259/0007-1285-46-552-1016>

3. Röntgen WC (1896) On a New Kind of Rays. *Science* 3(59):227–231
4. Ter-Pogossian MM, Phelps ME, Hoffman EJ, Mullani NA (1975) A Positron-Emission Transaxial Tomograph for Nuclear Imaging (PETT). *Radiology* 114(1):89–98. <https://doi.org/10.1148/1141.189>
5. Jaszczak RJ, Murphy PH, Huard D, Burdine JA (1977) Radionuclide emission computed tomography of the head with ^{99m}Tc and a Scintillation Camera. *J Nucl Med* 18(4): 373–380
6. Keyes JW, Orlandea N, Heetderks WJ, Leonard PF, Rogers WL (1977) The Humongotron—A Scintillation-Camera Transaxial Tomograph. *J Nucl Med* 18(4): 381–387
7. Becquerel H (1903) Recherches Sur Une Propriété Nouvelle de La Matière: Activité Radiante Spontanée Ou Radioactivité de La Matière. *Mémoires de l'Académie Des Sciences de l'Institut de France, L'Institut de France*
8. Young IR, Bailes DR, Burl M, Collins AG, Smith DT, McDonnell MJ, Orr JS, Banks LM, Bydder GM, Greenspan RH, Steiner RE (1982) Initial clinical evaluation of a whole body nuclear magnetic resonance (NMR) Tomograph. *J Comput Assist Tomogr* 6(1):1–18. <https://doi.org/10.1097/00004728-198202000-00001>
9. Bloch F (1946) Nuclear induction. *Phys Rev* 70(7–8):460–474. <https://doi.org/10.1103/PhysRev.70.460>
10. Townsend DW, Beyer T, Blodgett TM (2003) PET/CT scanners: A hardware approach to image fusion. *Semin Nucl Med* 33(3): 193–204. <https://doi.org/10.1053/snuc.2003.127314>
11. Schlemmer HP, Pichler B, Wienhard K, Schmand M, Nahmias C, Townsend D, Heiss WD, Claussen C (2007) Simultaneous MR/PET for brain imaging: First patient scans. *J Nucl Med* 48(supplement 2):45P–45P
12. Schmand M, Burbar Z, Corbeil J, Zhang N, Michael C, Byars L, Eriksson L, Grazioso R, Martin M, Moor A, Camp J, Matschl V, Ladbeck R, Renz W, Fischer H, Jattke K, Schnur G, Rietsch N, Bendriem B, Heiss WD (2007) BrainPET: First human tomograph for simultaneous (functional) PET and MR imaging. *J Nucl Med* 48(supplement 2): 45P–45P
13. Patton JA, Delbeke D, Sandler MP (2000) Image fusion using an integrated, dual-head coincidence camera with X-ray tube-based attenuation maps. *J Nucl Med* 41(8): 1364–1368
14. Hutton BF, Occhipinti M, Kuehne A, Máthé D, Kovács N, Waiczies H, Erlandsson K, Salvado D, Carminati M, Montagnani GL et al (2018) Development of clinical simultaneous SPECT/MRI. *Br J Radiol* 91(1081):20160690. <https://doi.org/10.1259/bjr.20160690>
15. Smith-Bindman R, Kwan ML, Marlow EC, Theis MK, Bolch W, Cheng SY, Bowles EJA, Duncan JR, Greenlee RT, Kushi LH, Pole JD, Rahm AK, Stout NK, Weinmann S, Miglioretti DL (2019) Trends in use of medical imaging in US Health Care Systems and in Ontario, Canada, 2000–2016. *JAMA* 322(9):843–856. <https://doi.org/10.1001/jama.2019.11456>
16. Evans AC, Janke AL, Collins DL, Baillet S (2012) Brain templates and atlases. *Neuroimage* 62(2):911–922. <https://doi.org/10.1016/j.neuroimage.2012.01.024>
17. Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage* 9(2):179–194. <https://doi.org/10.1006/nimg.1998.0395>
18. Fischl B, Sereno MI, Dale AM (1999) Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. *Neuroimage* 9(2):195–207. <https://doi.org/10.1006/nimg.1998.0396>
19. de Vico Fallani F, Richiardi J, Chavez M, Achard S (2014) Graph analysis of functional brain networks: practical issues in translational neuroscience. *Philos Trans R Soc B Biol Sci* 369(1653):20130521. <https://doi.org/10.1098/rstb.2013.0521>
20. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, Flandin G, Ghosh SS, Glatard T, Halchenko YO, Handwerker DA, Hanke M, Keator D, Li X, Michael Z, Maumet C, Nichols BN, Nichols TE, Pellman J, Poline JB, Rokem A, Schaefer G, Sochat V, Triplett W, Turner JA, Varoquaux G, Poldrack RA (2016) The brain imaging data structure, a format for organizing and describing outputs of neuro-imaging experiments. *Sci Data* 3: sdata201644. <https://doi.org/10.1038/sdata.2016.44>
21. Penny WD, Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE (2011) Statistical parametric mapping: the analysis of functional brain images. Elsevier, Amsterdam
22. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM (2012) FSL.

- NeuroImage 62(2):782–790. <https://doi.org/10.1016/j.neuroimage.2011.09.015>
23. Fischl B (2012) Freesurfer. NeuroImage 62(2):774–781. <https://doi.org/10.1016/j.neuroimage.2012.01.021>
24. Avants BB, Tustison N, Song G (2009) Advanced normalization tools (ANTS). The Insight Journal 2(365):1–35. <https://doi.org/10.54294/uvnhin>
25. Tournier JD, Calamante F, Connelly A (2012) MRtrix: diffusion tractography in crossing fiber regions. Int J Imaging Syst Technol 22(1):53–66. <https://doi.org/10.1002/ima.22005>
26. Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29(3):162–173. <https://doi.org/10.1006/cbmr.1996.0014>
27. Gorgolewski KJ, Alfaro-Almagro F, Auer T, Bellec P, Capotă M, Chakravarty MM, Churchill NW, Cohen AL, Craddock RC, Devenyi GA, Eklund A, Esteban O, Flandin G, Ghosh SS, Guntupalli JS, Jenkinson M, Keshavan A, Kiar G, Liem F, Raamana PR, Raffelt D, Steele CJ, Quirion PO, Smith RE, Strother SC, Varoquaux G, Wang Y, Yarkoni T, Poldrack RA (2017) BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. PLoS Comput Biol 13(3): 1–16. <https://doi.org/10.1371/journal.pcbi.1005209>
28. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, Gramfort A, Thirion B, Varoquaux G (2014) Machine learning for neuroimaging with scikit-learn. Front Neuroinform 8:14. <https://doi.org/10.3389/fninf.2014.00014>
29. Routier A, Burgos N, Díaz M, Bacci M, Bottani S, El-Rifai O, Fontanella S, Gori P, Guillón J, Guyot A, Hassanal R, Jacquemont T, Lu P, Marcoux A, Moreau T, Samper-González J, Teichmann M, Thibreau-Sutre E, Vaillant G, Wen J, Wild A, Habert MO, Durrleman S, Colliot O (2021) Clinica: an open-source software platform for reproducible clinical neuroscience studies. Front Neuroinform 15:39. <https://doi.org/10.3389/fninf.2021.689675>
30. Pérez-García F, Sparks R, Ourselin S (2021) TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput Methods Prog Biomed 208:106236. <https://doi.org/10.1016/j.cmpb.2021.106236>
31. Thibreau-Sutre E, Díaz M, Hassanal R, Routier A, Dormont D, Colliot O, Burgos N (2022) ClinicaDL: An open-source deep learning software for reproducible neuroimaging processing. Comput Methods Prog Biomed 220:106818. <https://doi.org/10.1016/j.cmpb.2022.106818>
32. Commowick O, Istace A, Kain M, Laurent B, Leray F, Simon M, Pop SC, Girard P, Améli R, Ferré JC, Kerbrat A, Tourdias T, Cervenansky F, Glataud T, Beaumont J, Doyle S, Forbes F, Knight J, Khademi A, Mahbod A, Wang C, McKinley R, Wagner F, Muschelli J, Sweeney E, Roura E, Lladó X, Santos MM, Santos WP, Silva-Filho AG, Tomas-Fernandez X, Urien H, Bloch I, Valverde S, Cabezas M, Vera-Olmos FJ, Malpica N, Guttmann C, Vukusic S, Edan G, Dojat M, Styner M, Warfield SK, Cotton F, Barillot C (2018) Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Sci Rep 8(1):13650. <https://doi.org/10.1038/s41598-018-31911-7>
33. Commowick O, Kain M, Casey R, Ameli R, Ferré JC, Kerbrat A, Tourdias T, Cervenansky F, Camarasu-Pop S, Glataud T, Vukusic S, Edan G, Barillot C, Dojat M, Cotton F (2021) Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset. NeuroImage 244:118589. <https://doi.org/10.1016/j.neuroimage.2021.118589>
34. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC (2010) N4ITK: Improved N3 Bias Correction. IEEE Trans Med Imaging 29(6): 1310–1320. <https://doi.org/10.1109/TMI.2010.2046908>
35. Nyúl LG, Udupa JK, Zhang X (2000) New variants of a method of MRI scale standardization. IEEE Trans Med Imaging 19(2): 143–150. <https://doi.org/10.1109/42.836373>
36. Kalavathi P, Prasath V (2016) Methods on skull stripping of MRI head scan images—a review. J Digit Imaging 29(3):365–379. <https://doi.org/10.1007/s10278-015-9847-8>
37. Shattuck DW, Leahy RM (2002) BrainSuite: an automated cortical surface identification tool. Med Image Anal 6(2):129–142. [https://doi.org/10.1016/S1361-8415\(02\)00054-3](https://doi.org/10.1016/S1361-8415(02)00054-3)
38. Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, Wick A, Schlemmer HP, Heiland S, Wick W et al (2019)

- Automated brain extraction of multisequence MRI using artificial neural networks. *Hum Brain Mapp* 40(17):4952–4964. <https://doi.org/10.1002/hbm.24750>
39. Hoopes A, Mora JS, Dalca AV, Fischl B, Hoffmann M (2022) SynthStrip: skull-stripping for any brain image. *NeuroImage* 260: 119474. <https://doi.org/10.1016/j.neuroimage.2022.119474>
 40. Oliveira FP, Tavares JMR (2014) Medical image registration: a review. *Comput Meth Biomed Eng* 17(2):73–93. <https://doi.org/10.1080/10255842.2012.670855>
 41. Dora L, Agrawal S, Panda R, Abraham A (2017) State-of-the-art methods for brain tissue segmentation: A review. *IEEE Rev Biomed Eng* 10:235–249. <https://doi.org/10.1109/RBME.2017.2715350>
 42. González-Villà S, Oliver A, Valverde S, Wang L, Zwiggelaar R, Lladó X (2016) A review on brain structures segmentation in magnetic resonance imaging. *Artif Intell Med* 73:45–69. <https://doi.org/10.1016/j.artmed.2016.09.001>
 43. Ashburner J, Friston KJ (2000) Voxel-Based Morphometry—The Methods. *NeuroImage* 11(6):805–821. <https://doi.org/10.1006/nimg.2000.0582>
 44. Fischl B, Dale AM (2000) Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc Natl Acad Sci* 97(20):11050–11055. <https://doi.org/10.1073/pnas.200033797>
 45. He Y, Chen ZJ, Evans AC (2007) Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. *Cereb Cortex* 17(10):2407–2419. <https://doi.org/10.1093/cercor/bhl149>
 46. Klöppel S, Stonnington CM, Chu C, Draganski B, Scähill RI, Rohrer JD, Fox NC, Jack Jr CR, Ashburner J, Frackowiak RS (2008) Automatic classification of MR scans in Alzheimer’s disease. *Brain* 131(3): 681–689. <https://doi.org/10.1093/brain/awm319>
 47. Rathore S, Habes M, Iftikhar MA, Shacklett A, Davatzikos C (2017) A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer’s disease and its prodromal stages. *NeuroImage* 155:530–548. <https://doi.org/10.1016/j.neuroimage.2017.03.057>
 48. Samper-González J, Burgos N, Bottani S, Fontanella S, Lu P, Marcoux A, Routier A, Guillou J, Bacci M, Wen J, Bertrand A, Bertin H, Habert MO, Durrleman S, Evgeniou T, Colliot O (2018) Reproducible evaluation of classification methods in Alzheimer’s disease: Framework and application to MRI and PET data. *NeuroImage* 183:504–521. <https://doi.org/10.1016/j.neuroimage.2018.08.042>
 49. Wen J, Thibaut-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Durrleman S, Burgos N, Colliot O (2020) Convolutional neural networks for classification of Alzheimer’s disease: Overview and reproducible evaluation. *Med Image Anal* 63:101694. <https://doi.org/10.1016/j.media.2020.101694>
 50. Bakas S, Reyes M, Jakab A, ..., Davatzikos C, van Leemput K, Menze B (2018) Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:181102629*
 51. Zeineldin RA, Karar ME, Coburger J, Wirtz CR, Burgert O (2020) DeepSeg: Deep neural network framework for automatic brain tumor segmentation using magnetic resonance FLAIR images. *Int J Comput Assist Radiol Surg* 15(6):909–920. <https://doi.org/10.1007/s11548-020-02186-z>
 52. Sweeney EM, Vogelstein JT, Cuzzocreo JL, Calabresi PA, Reich DS, Crainiceanu CM, Shinohara RT (2014) A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structural MRI. *PLoS One* 9(4): e95753. <https://doi.org/10.1371/journal.pone.0095753>
 53. La Rosa F, Abdulkadir A, Fartaria MJ, Rahmazadeh R, Lu P, Galbusera R, Barakovic M, Thiran J, Granziera C, Cuadra MB (2020) Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: A deep learning method based on FLAIR and MP2RAGE. *NeuroImage: Clinical* 27: 102335. <https://doi.org/10.1016/j.nicl.2020.102335>
 54. Le Bihan D, Breton E, Lallemand D, Grenier P, Cabanis E, Laval-Jeantet M (1986) MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. *Radiology* 161(2):401–407
 55. Tournier JD (2019) Diffusion MRI in the brain—Theory and concepts. *Prog Nucl Magn Reson Spectrosc* 112:1–16. <https://doi.org/10.1016/j.pnmr.2019.03.001>
 56. Tax CM, Bastiani M, Veraart J, Garyfallidis E, Irfanoglu MO (2022) What’s new and what’s

- next in diffusion MRI preprocessing. NeuroImage 249:118830. <https://doi.org/10.1016/j.neuroimage.2021.118830>
57. Soares JM, Marques P, Alves V, Sousa N (2013) A hitchhiker's guide to diffusion tensor imaging. *Front. Neurosci.* 7:31. <https://doi.org/10.3389/fnins.2013.00031>
 58. Jeurissen B, Descoteaux M, Mori S, Leemans A (2019) Diffusion MRI fiber tractography of the brain. *NMR Biomed* 32(4):e3785. <https://doi.org/10.1002/nbm.3785>
 59. Zhang H, Schneider T, Wheeler-Kingshott CA, Alexander DC (2012) NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *NeuroImage* 61(4):1000–1016. <https://doi.org/10.1016/j.neuroimage.2012.03.072>
 60. Maggipinto T, Bellotti R, Amoroso N, Diacono D, Donvito G, Lella E, Monaco A, Scelsi MA, Tangaro S (2017) DTI measurements for Alzheimer's classification. *Phys Med Biol* 62(6):2361. <https://doi.org/10.1088/1361-6560/aa5dbe>
 61. Wen J, Samper-González J, Bottani S, Routier A, Burgos N, Jacquemont T, Fontanella S, Durrleman S, Epelbaum S, Bertrand A, Colliot O (2021) Reproducible evaluation of diffusion MRI features for automatic classification of patients with Alzheimer's disease. *Neuroinformatics* 19(1): 57–78. <https://doi.org/10.1007/s12021-020-09469-5>
 62. Park YW, Oh J, You SC, Han K, Ahn SS, Choi YS, Chang JH, Kim SH, Lee SK (2019) Radiomics and machine learning may accurately predict the grade and histological subtype in meningiomas using conventional and diffusion tensor imaging. *Eur Radiol* 29(8): 4068–4076. <https://doi.org/10.1007/s00330-018-5830-3>
 63. Glover GH (2011) Overview of functional magnetic resonance imaging. *Neurosurg Clin* 22(2):133–139. <https://doi.org/10.1016/j.nec.2010.11.001>
 64. Fox MD, Raichle ME (2007) Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat Rev Neurosci* 8(9):700–711. <https://doi.org/10.1038/nrn2201>
 65. Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, Kent JD, Goncalves M, DuPre E, Snyder M, Oya H, Ghosh SS, Wright J, Durnez J, Poldrack RA, Gorgolewski KJ (2019) fMRIprep: A robust preprocessing pipeline for functional MRI. *Nat Methods* 16(1):111–116. <https://doi.org/10.1038/s41592-018-0235-4>
 66. Khosla M, Jamison K, Ngo GH, Kuceyeski A, Sabuncu MR (2019) Machine learning in resting-state fMRI analysis. *Magn Reson Imaging* 64:101–121. <https://doi.org/10.1016/j.mri.2019.05.031>
 67. Fischl B, Sereno MI, Tootell RB, Dale AM (1999) High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum Brain Mapp* 8(4):272–284. [https://doi.org/10.1002/\(SICI\)1097-0193\(1999\)8:4%3C272::AID-HBM10%3E3.0.CO;2-4](https://doi.org/10.1002/(SICI)1097-0193(1999)8:4%3C272::AID-HBM10%3E3.0.CO;2-4)
 68. Kim J, Calhoun VD, Shim E, Lee JH (2016) Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *NeuroImage* 124:127–146. <https://doi.org/10.1016/j.neuroimage.2015.05.018>
 69. Rashid B, Arbabshirani MR, Damaraju E, Cetin MS, Miller R, Pearlson GD, Calhoun VD (2016) Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity. *NeuroImage* 134:645–657. <https://doi.org/10.1016/j.neuroimage.2016.04.051>
 70. Wannamaker R, Buck B, Butcher K (2019) Multimodal CT in Acute Stroke. *Curr Neurol Neurosci Rep* 19(9):63. <https://doi.org/10.1007/s11910-019-0978-z>
 71. Muschelli J (2019) Recommendations for processing head CT data. *Front Neuroinform* 13:61. <https://doi.org/10.3389/fninf.2019.00061>
 72. Chourmouzi D, Papadopoulou E, Marias K, Drevelegas A (2014) Imaging of Brain Tumors. *Surg Oncol Clin* 23(4):629–684. <https://doi.org/10.1016/j.soc.2014.07.004>
 73. Yeo M, Tahayori B, Kok HK, Maingard J, Kutaiba N, Russell J, Thijss V, Jhamb A, Chandra RV, Brooks M, Barras CD, Asadi H (2021) Review of deep learning algorithms for the automatic detection of intracranial hemorrhages on computed tomography head imaging. *Journal of NeuroInterventional Surgery* 13(4):369–378. <https://doi.org/10.1136/neurintsurg-2020-017099>
 74. Buchlak QD, Milne MR, Seah J, Johnson A, Samarasinghe G, Hachey B, Esmaili N, Tran A, Leveque JC, Farrokhi F, Goldschlager T, Edelstein S, Brotchie P (2022) Charting the potential of brain computed tomography deep learning systems. *J*

- Clin Neurosci 99:217–223. <https://doi.org/10.1016/j.jocn.2022.03.014>
75. Ye H, Gao F, Yin Y, Guo D, Zhao P, Lu Y, Wang X, Bai J, Cao K, Song Q, Zhang H, Chen W, Guo X, Xia J (2019) Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. Eur Radiol 29(11):6191–6201. <https://doi.org/10.1007/s00330-019-06163-2>
 76. Clerigues A, Valverde S, Bernal J, Freixenet J, Oliver A, Lladó X (2019) Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks. Comput Biol Med 115:103487. <https://doi.org/10.1016/j.combiomed.2019.103487>
 77. Hooker JM, Carson RE (2019) Human positron emission tomography neuroimaging. Annu Rev Biomed Eng 21:551–581. <https://doi.org/10.1146/annurev-bioeng-062117-121056>
 78. Heurling K, Leuzy A, Jonasson M, Frick A, Zimmer ER, Nordberg A, Lubberink M (2017) Quantitative positron emission tomography in brain research. Brain Res 1670:220–234. <https://doi.org/10.1016/j.brainres.2017.06.022>
 79. Guedj E, Varrone A, Boellaard R, Albert NL, Barthel H, van Berckel B, Brendel M, Cecchin D, Ekmekcioglu O, Garibotto V, Lammertsma AA, Law I, Peñuelas I, Semah F, Traub-Weidinger T, van de Giessen E, Van Weehaeghe D, Morbelli S (2022) EANM procedure guidelines for brain PET imaging using [18F]FDG, version 3. Eur J Nucl Med Mol Imaging 49(2): 632–651. <https://doi.org/10.1007/s00259-021-05603-w>
 80. Cho H, Choi JY, Hwang MS, Kim YJ, Lee HM, Lee HS, Lee JH, Ryu YH, Lee MS, Lyoo CH (2016) In vivo cortical spreading pattern of tau and amyloid in the Alzheimer disease spectrum. Ann Neurol 80(2): 247–258. <https://doi.org/10.1002/ana.24711>
 81. Stankoff B, Freeman L, Aigrot MS, Chardain A, Dollé F, Williams A, Galanaud D, Armand L, Lehericy S, Lubetzki C et al (2011) Imaging central nervous system myelin by positron emission tomography in multiple sclerosis using [methyl-11C]-2-(4'-methylaminophenyl)-6-hydroxybenzothiazole. Ann Neurol 69(4):673–680. <https://doi.org/10.1002/ana.22320>
 82. Galldiks N, Lohmann P, Albert NL, Tonn JC, Langen KJ (2019) Current status of pet imaging in neuro-oncology. Neuro-Oncology Advances 1(1):vdz010. <https://doi.org/10.1093/noajnl/vdz010>
 83. Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, Harvey D, Jack CR, Jagust W, Morris JC, Petersen RC, Salazar J, Saykin AJ, Shaw LM, Toga AW, Trojanowski JQ (2017) The Alzheimer’s Disease Neuro-imaging Initiative 3: Continued innovation for clinical trial improvement. Alzheimers Dement. 13(5):561–571. <https://doi.org/10.1016/j.jalz.2016.10.006>
 84. López-González FJ, Silva-Rodríguez J, Paredes-Pacheco J, Niñerola-Baizán A, Efthimiou N, Martín-Martín C, Moscoso A, Ruibal A, Roé-Vellvé N, Aguiar P (2020) Intensity normalization methods in brain fdg-pet quantification. Neuroimage 222: 117229. <https://doi.org/10.1016/j.neuroimage.2020.117229>
 85. Erlandsson K, Buvat I, Pretorius PH, Thomas BA, Hutton BF (2012) A review of partial volume correction techniques for emission tomography and their applications in neurology, cardiology and oncology. Phys Med Biol 57(21):R119. <https://doi.org/10.1088/0031-9155/57/21/R119>
 86. Thomas BA, Cuplov V, Bousse A, Mendes A, Thielemans K, Hutton BF, Erlandsson K (2016) PETPVC: A toolbox for performing partial volume correction techniques in positron emission tomography. Phys Med Biol 61(22):7975–7993. <https://doi.org/10.1088/0031-9155/61/22/7975>
 87. Della Rosa PA, Cerami C, Gallivanone F, Prestia A, Caroli A, Castiglioni I, Gilardi MC, Frisoni G, Friston K, Ashburner J, Perani D (2014) A Standardized 18F-FDG-PET Template for spatial normalization in statistical parametric mapping of Dementia. Neuroinformatics 12(4):575–593. <https://doi.org/10.1007/s12021-014-9235-4>
 88. Marcoux A, Burgos N, Bertrand A, Teichmann M, Routier A, Wen J, Samper-González J, Bottani S, Durrleman S, Habert MO, Colliot O (2018) An automated pipeline for the analysis of PET Data on the cortical surface. Front Neuroinform 12:94. <https://doi.org/10.3389/fninf.2018.00094>
 89. Yakushev I, Drzezga A, Habeck C (2017) Metabolic connectivity: methods and applications. Curr Opin Neurol 30(6):677–685. <https://doi.org/10.1097/WCO.0000000000000494>
 90. Duffy IR, Boyle AJ, Vasdev N (2019) Improving PET imaging acquisition and analysis with machine learning: a narrative review with focus on Alzheimer’s disease and oncology. Mol Imaging 18:1536012119869070.

- <https://doi.org/10.1177/1536012119869070>
91. Gray KR, Wolz R, Heckemann RA, Aljabar P, Hammers A, Rueckert D, Initiative ADN et al (2012) Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease. *NeuroImage* 60(1):221–229. <https://doi.org/10.1016/j.neuroimage.2011.12.071>
 92. Lu D, Popuri K, Ding GW, Balachandar R, Beg MF (2018) Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease. *Med Image Anal* 46:26–34. <https://doi.org/10.1016/j.media.2018.02.002>
 93. Higdon R, Foster NL, Koeppe RA, DeCarli CS, Jagust WJ, Clark CM, Barbas NR, Arnold SE, Turner RS, Heidebrink JL, Minoshima S (2004) A comparison of classification methods for differentiating fronto-temporal dementia from Alzheimer's disease using FDG-PET imaging. *Stat Med* 23(2):315–326. <https://doi.org/10.1002/sim.1719>
 94. Papp L, Pötsch N, Grahovac M, Schmidbauer V, Woehrer A, Preusser M, Mitterhauser M, Kiesel B, Wadsak W, Beyer T et al (2018) Glioma survival prediction with combined analysis of in vivo ¹¹C-MET PET features, ex vivo features, and patient features by supervised machine learning. *J Nucl Med* 59(6):892–899. <https://doi.org/10.2967/jnumed.117.202267>
 95. Hotta M, Minamimoto R, Miwa K (2019) ¹¹C-methionine-PET for differentiating recurrent brain tumor from radiation necrosis: radiomics approach with random forest classifier. *Sci Rep* 9(1):1–7. <https://doi.org/10.1038/s41598-019-52279-2>
 96. Accorsi R (2008) Brain Single-photon emission CT physics principles. *Am J Neuroradiol* 29(7):1247–1256. <https://doi.org/10.3174/ajnr.A1175>
 97. Kapucu ÖL, Nobili F, Varrone A, Booij J, Vander Borght T, Någren K, Darcourt J, Tatsch K, Van Laere KJ (2009) EANM procedure guideline for brain perfusion SPECT using ^{99m}Tc-labelled radiopharmaceuticals, version 2. *Eur J Nucl Med Mol Imaging* 36(12):2093. <https://doi.org/10.1007/s00259-009-1266-y>
 98. Morbelli S, Esposito G, Arbizu J, Barthel H, Boellaard R, Bohnen NI, Brooks DJ, Darcourt J, Dickson JC, Douglas D, Drzezga A, Dubroff J, Ekmekcioglu O, Garibotti V, Herscovitch P, Kuo P, Lammertsma A, Pappata S, Peñuelas I, Seibyl J, Semah F, Tossici-Bolt L, Van de Giessen E, Van Laere K, Varrone A, Wanner M, Zubal G, Law I (2020) EANM practice guideline/SNMMI procedure standard for dopaminergic imaging in Parkinsonian syndromes 1.0. *Eur J Nucl Med Mol Imaging* 47(8):1885–1912. <https://doi.org/10.1007/s00259-020-04817-8>
 99. Yeo JM, Lim X, Khan Z, Pal S (2013) Systematic review of the diagnostic utility of SPECT imaging in dementia. *Eur Arch Psychiatry Clin Neurosci* 263(7):539–552. <https://doi.org/10.1007/s00406-013-0426-z>
 100. McNally KA, Paige AL, Varghese G, Zhang H, Novotny Jr EJ, Spencer SS, Zubal IG, Blumenfeld H (2005) Localizing value of ictal–interictal SPECT analyzed by SPM (ISAS). *Epilepsia* 46(9):1450–1464. <https://doi.org/10.1111/j.1528-1167.2005.06705.x>
 101. Marek K, Chowdhury S, Siderowf A, Lasch S, Coffey CS, Caspell-Garcia C, Simuni T, Jennings D, Tanner CM, Trojanowski JQ, Shaw LM, Seibyl J, Schuff N, Singleton A, Kieburtz K, Toga AW, Mollenhauer B, Galasko D, Chahine LM, Weintraub D, Foroud T, Tosun-Turgut D, Poston K, Arnedo V, Frasier M, Sherer T (2018) The Parkinson's progression markers initiative (PPMI)—establishing a PD biomarker cohort. *Ann Clin Transl Neurol* 5(12):1460–1477. <https://doi.org/10.1002/acn3.644>
 102. Khachnaoui H, Mabrouk R, Khelifa N (2020) Machine learning and deep learning for clinical data and PET/SPECT imaging in Parkinson's disease: A review. *IET Image Process* 14(16):4013–4026. <https://doi.org/10.1049/iet-ipr.2020.1048>
 103. Prashanth R, Roy SD, Mandal PK, Ghosh S (2014) Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging. *Expert Systems with Applications* 41(7):3333–3342. <https://doi.org/10.1016/j.eswa.2013.11.031>
 104. Choi H, Ha S, Im HJ, Paek SH, Lee DS (2017) Refining diagnosis of parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. *NeuroImage: Clinical* 16:586–594. <https://doi.org/10.1016/j.nicl.2017.09.010>
 105. Rahmim A, Huang P, Shenkov N, Fotouhi S, Davoodi-Bojd E, Lu L, Mari Z, Soltanian-Zadeh H, Sossi V (2017) Improved prediction of outcome in Parkinson's disease using radiomics analysis of longitudinal DAT SPECT images. *NeuroImage: Clinical* 16:539–544. <https://doi.org/10.1016/j.nicl.2017.08.021>

106. Fung G, Stoeckel J (2007) SVM feature selection for classification of spect images of alzheimer's disease using spatial information. *Knowl Inf Syst* 11(2):243–258. <https://doi.org/10.1007/s10115-006-0043-5>
107. Górriz J, Segovia F, Ramírez J, Lassl A, Salas-Gonzalez D (2011) GMM based SPECT image classification for the diagnosis of Alzheimer's disease. *Appl Soft Comput* 11(2): 2313–2325. <https://doi.org/10.1016/j.asoc.2010.08.012>
108. de Galiza Barbosa F, Delso G, Ter Voert E, Huellner M, Herrmann K, Veit-Haibach P (2016) Multi-technique hybrid imaging in PET/CT and PET/MR: what does the future hold? *Clin Radiol* 71(7):660–672. <https://doi.org/10.1016/j.crad.2016.03.013>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 9

Electroencephalography and Magnetoencephalography

Marie-Constance Corsi

Abstract

In this chapter, we present the main characteristics of electroencephalography (EEG) and magnetoencephalography (MEG). More specifically, this chapter is dedicated to the presentation of the data, the way they can be acquired and analyzed. Then, we present the main features that can be extracted and their applications for brain disorders with concrete examples to illustrate them. Additional materials associated with this chapter are available in the dedicated [Github repository](#).

Key words Electroencephalography, Magnetoencephalography, Evoked activity, Oscillatory activity, Brain-computer interfaces

1 Introduction

This chapter aims at providing an overview of electroencephalography (EEG) and magnetoencephalography (MEG) to help the reader with no previous experience with these modalities to understand the information that can be extracted and their neurophysiological meaning in the perspective to be used for brain disorders. These two modalities, which share common characteristics, are often designated together with the acronym M/EEG.

To this end, instead of providing an exhaustive presentation of the M/EEG clinical applications, we focused on the main aspects related to these modalities. As a result, this chapter is organized as follows: We first describe the basic principles in terms of origins of the signals and electrophysiological activity exploited in M/EEG (Subheading 2). We then present the principles of M/EEG experiments (Subheading 3), the data analysis techniques (Subheading 4), and in particular features that can be extracted from the data (Subheading 5). The last part of this chapter presents illustrations of M/EEG applications to brain disorders (Subheading 6). To go further, additional resources are provided to the reader in Boxes 1 and 2 and in a dedicated [Github repository](#).

2 Basic Principles

Being able to extract the information of interest to perform a classification from M/EEG data requires to have some neurophysiological background knowledge to assess the relevance of the selected features. This paragraph aims at providing some general elements regarding the origin of the signals and the recorded activity.

2.1 Origin of the Signals

Neurons create electrical signals, transmitted to other cells via synapses. First, an action potential (AP) arrives at a synaptic cleft (step 1 in Fig. 1) where it will transmit chemical information via neurotransmitters (step 2 in Fig. 1) that generate postsynaptic potentials (PSPs) and local currents (step 3 in Fig. 1). A PSP will create a current sink and will propagate until the cell body to generate a current source (step 4 in Fig. 1). As a result, the PSP creates an electrical dipole consisting in a negative pole (i.e., the sink) and a positive pole (i.e., the source). This dipole will generate primary (intracellular) currents and secondary (extracellular) currents. M/EEG signals result from postsynaptic potentials. More specifically, M/EEG signals result from the spatial and temporal summation of the activity of a large population of synchronous neurons. But notable differences exist between MEG and EEG.

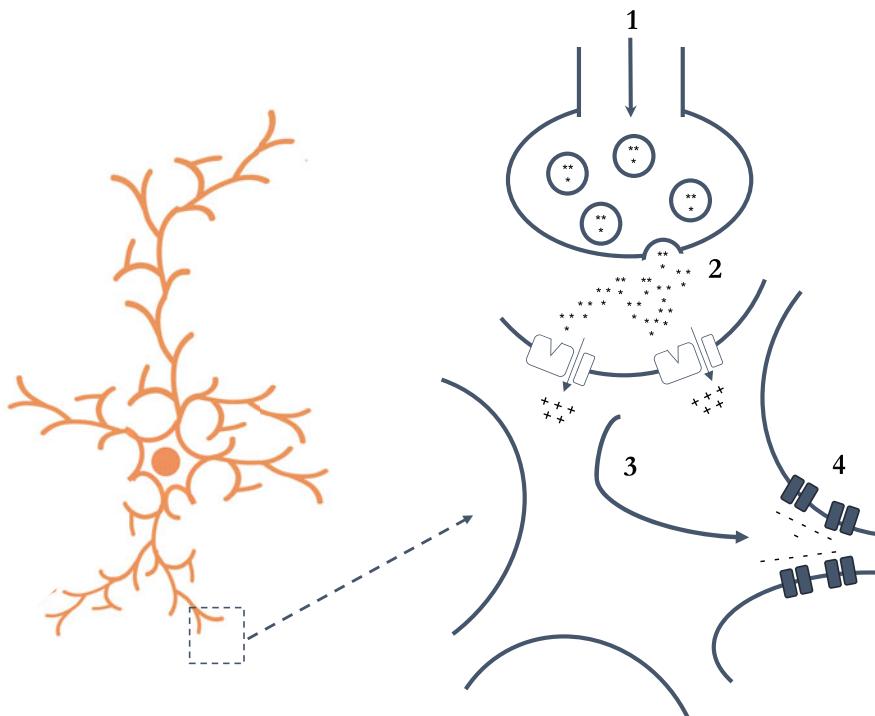


Fig. 1 Origin of M/EEG signals

Table 1
Main features to compare MEG and EEG

Items	MEG	EEG
Measurement	Magnetic field, + intracellular currents	Difference of potentials, + extracellular currents
Spatial resolution	1 cm	2–3 cm
Temporal resolution		1 ms or less
Amplitudes	≈ 100 fT	≈ 100 μ Volts
Advantages	<ul style="list-style-type: none"> – Absolute values – Less affected by bone – Focal – Focal 	<ul style="list-style-type: none"> – Portable – Cost
Drawbacks	<ul style="list-style-type: none"> – Financial constraints – Mechanical constraints 	<ul style="list-style-type: none"> – Need of a reference – Affected by bone – Diffuse

Firstly, regarding the signals themselves, MEG signals are mainly caused by intracellular currents generated by the PSP at the dendrite level and less by the extracellular currents; EEG signals correspond to a difference between electrical potentials, mainly due to extracellular currents. Secondly, regarding the sensitivity toward the dipole orientation, EEG is sensitive both to radial currents (activity located at the gyrus level) and to tangential currents (generated within sulci) even though it has stronger sensitivity to radial currents, whereas MEG is more sensitive to tangential currents. Finally, regarding the sensitivity toward the conductivity, EEG is strongly attenuated and deformed by crossing through the skull, whereas MEG is less sensitive to the different layers crossed (i.e., skull, brain, etc.). Such differences between MEG and EEG have an impact on the way data are preprocessed, analyzed, and, therefore, interpreted. The differences between MEG and EEG are summarized in Table 1.

2.2 Evoked and Oscillatory Activity

There are two main types of electrophysiological activity of interest that are exploited in the M/EEG domain: the evoked and the oscillatory activity. Evoked responses are weak variations of electromagnetic activity resulting from a stimulation (for instance, in response to a task performance by the participant). Given their amplitude, it is often necessary to average signals over chunks of signals, referred as epochs, to reduce noise. To identify and describe these evoked responses, there is a specific way to name them according to their latency, their amplitude, their shape, and the

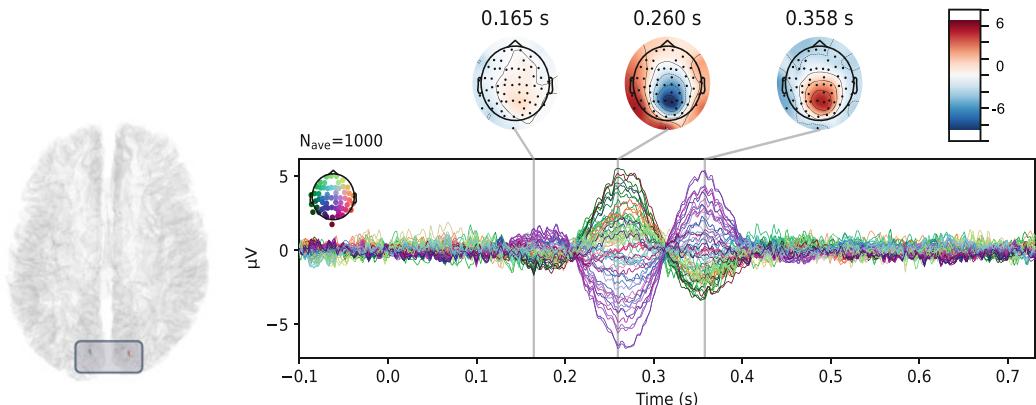


Fig. 2 Evoked activity. Results from a simulation where two sources, located in the visual area, generated an activity after a stimulus. On the right, we plotted the associated time course over the scalp (synthetic signals), resulting from the averaging of 1000 repetitions. One can observe notably a positive wave around $t = 300$ ms. The code to generate this figure is accessible via the dedicated [Github repository](#)

polarity. Let's take an example (see Fig. 2), which represent evoked responses from a study where we simulated a visual stimulation. We first see a positive deflection occurring 300 ms after the presentation of the stimulation, which is referred to as P300. These waves can reflect different mechanisms: the early components are mostly exogenous and are related to the stimulus characteristics; the late components are endogenous and are related to the performed task and to the subject's state.

The oscillatory activity, or induced activity, results from the summation of the activity in a given brain region. These rhythms are mainly defined by their frequency, their amplitude, their shape, their location, and their duration. In Fig. 3, we provided examples of the main rhythms found in the literature. Each frequency band is referred to by a Greek letter. Delta ([0.5–3 Hz]) and Theta ([3–7 Hz]) rhythms are, respectively, detected in deep and slight sleeps. Alpha ([8–12 Hz] in posterior areas) and Mu ([7–13 Hz] in central areas) rhythms are both observed in quiet watch and resting state (with the eyes closed for Alpha). Beta ([13–30 Hz]) rhythm is detected during the active watch and during cognitive tasks such as motor imagery, for instance. Gamma rhythm (divided into two sub-rhythms: slow in 30–70 Hz and fast beyond 70 Hz) is observed during specific cognitive processing.

3 M/EEG Experiments

This section provides an overview of the devices currently used and the main steps that constitute an M/EEG experiment. As a take-home message, in Table 1, we propose a comparison of the main features of MEG and EEG.

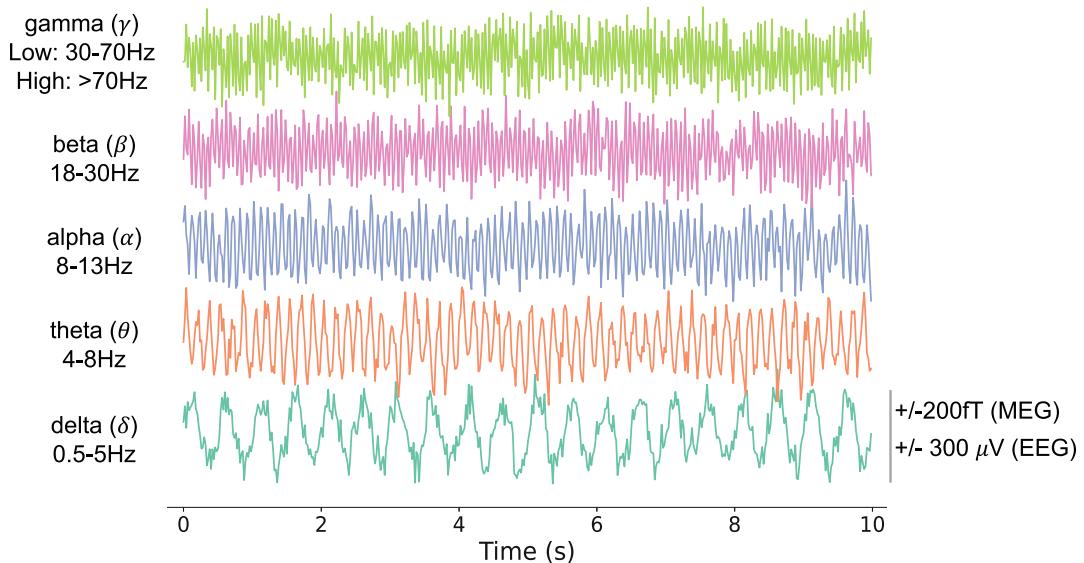


Fig. 3 We plotted the time course associated with the main rhythms that one can observe from M/EEG recordings. These plots were obtained from synthetic signals. The code to generate this figure is accessible via the dedicated [Github repository](#)

3.1 Instrumentation

3.1.1 EEG

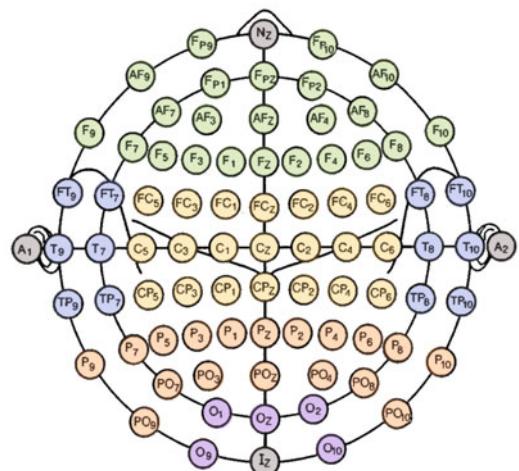
EEG signals are recorded through the use of electrodes placed over the scalp. The EEG relies on the difference of potentials. The first EEG recordings have been performed by Hans Berger in 1924. He described the oscillatory activity at 8 Hz occurring in the posterior area of the scalp when the subject is awake with his eyes closed. There are different types of electrodes: wet/dry electrodes and active/passive electrodes. Wet electrodes are generally made of tin, silver, or silver chloride material (Ag/AgCl). They need an electrolytic gel to enable the conduction between the skin and the electrode. Dry electrodes are made of stainless steel that behaves as a conductor between the skin and the electrode. The active electrodes contain an electronic module that performs a pre-amplification of the signal to ensure the stability of the system toward changes in impedance and noise. The passive electrodes do not use a pre-amplification module.

Naming Even though some differences may be found from one EEG device to another, there are some standardized ways to name and localize EEG sensors (also called channels). Each channel is often referred to by a letter and a number. Most of the time, odd channels are located on the left hemisphere and the even ones on the right hemisphere. The letters correspond to the area: frontal, temporal, parietal, central, and occipital. In addition to the sensors themselves, one can also find landmarks: nasion, inion, and pre-auricular points. An example of such naming is shown in Fig. 4b.

A



B



C



Fig. 4 M/EEG instrumentation. **(a)** EEG experimental setup. **(b)** Example of EEG montage. For an illustrative purpose, each color corresponds to a brain area. Each circle represents either a sensor or a landmark. Sensors appear in color while landmarks appear in gray. Sensors are designated with a letter and a number. The letter is indicative of the brain region. Odd numbers correspond to the left hemisphere and even ones to the right hemisphere. **(c)** MEG experimental setup

List of Montages Depending on the scientific question to be addressed and, therefore, the brain areas of interest, different montages can be found. One can build an EEG montage from less than 5 electrodes to up to 256 channels. EEG measurements rely on a difference of electrical potentials. For this purpose, two montages can be considered: the referential montage and the bipolar montage. In the referential montage, each difference of electrical potentials considers an electrode placed over the scalp and a reference. As a result, each electrode placed over the scalp is compared to the reference electrode. The choice of the reference is crucial. The most commonly chosen locations for the reference electrodes are the mastoids (i.e., temporal bone behind the ears), even though several studies prefer placing the reference at the vertex (Cz, i.e., midline central of the scalp). Again, the location depends on the scientific question to be addressed. The bipolar montage consists of performing the difference between two electrodes placed over the scalp, after the experiment. Another electrode, referred to as ground electrode, is used. Among the privileged locations is the scapula (i.e., shoulder blade). An example of an EEG setup and a standard montage are proposed in Fig. 4a, b. For a complete description of the standardized EEG electrode arrays, the reader can refer to [1].

Future of EEG Hardware In the past years, there has been an increased interest in developing wearable EEG, to remove wires and to reduce its dimension but also to enable long-lasting recordings in a less constrained environment. Three bottlenecks need to be overcome: the EEG electrodes, hard to put on and to keep in place on the head; the EEG hardware, to make it less power-consuming and miniaturized; and the EEG software, to propose the most intelligible and reliable information regarding the captured brain activity [2]. In particular, EEG systems that rely on dry EEG electrodes get more and more attention. By not requiring conductive gel, it reduces the preparation time. Recent studies relying on commercialized dry electrodes systems show performances close to those obtained with wet electrodes [2].

3.1.2 MEG

Sensors and Main Devices The difficulty here is to detect signals that are 10^9 weaker than the Earth magnetic field. The current devices rely on superconducting quantum interference devices (SQUIDs) that can detect small MEG signals [3]. One of the first proof of concept was made by D. Cohen in the 1970s [4]. The SQUIDs present a sensitivity, defined here as the smallest variation of magnetic field that can be detected by the sensor, of $1 fT/\sqrt{Hz}$. To obtain such performance, a magnetic shielding room is required to remove the environmental noise, and a part of the device needs to be cooled via a cryogenic system (see Fig. 4c). Two types of

sensors are used to record MEG signals: magnetometers and gradiometers. Magnetometers measure the magnetic field, whereas the gradiometers measure the gradient of the magnetic field. They are used for noise elimination and consist in a combination of magnetometers. The main difference from one manufacturer to another lies in the type of gradiometers used:

- CTF manufacturer: radial gradiometers consisting of two magnetometers placed one above the other
- MEGIN manufacturer: planar gradiometers consisting of two magnetometers placed side by side

The type of gradiometer has an influence on the way brain activity is recorded and, therefore, on how to interpret the recorded signal [5]. Magnetometers and radial gradiometers are more sensitive to sources around the sensor, whereas planar gradiometers are more sensitive to sources located right below the sensor (Fig. 5).

New Generation of Sensors The current devices rely on a cryogenic cooling system that engenders technical and financial constraints. New cryogenic-free sensors have recently emerged: the optically pumped magnetometers (OPMs) [6, 7]. Developing cryogenic-free sensors presents two main advantages: an increase in the amplitude of the signal recorded by the sensor and a reduction of the dimension of the magnetic shielding room. Recent studies proved that OPMs present a better signal-to-noise ratio than EEG [8], can detect deep sources [9], and can be suited for pediatric or movement disorder studies [10]. Promising results could be obtained with triaxial measurements obtained from OPMs [11, 12].

3.2 Data Acquisition

Depending on the tasks and on the hardware used, the duration of an M/EEG experiment may vary. This section aims to present the main steps that constitute the data acquisition.

The first step consists in preparing all the materials to perform the experiment. For EEG, it will consist in cleaning the locations where electrodes will be in contact with the skin (e.g., forehead and mastoids). The electrodes and the EEG cap are then placed. Several key distances can be measured to verify that the cap is well-placed or to record fiducial points to be matched to other modalities afterward (e.g., MRI). Then, the experimenter needs to ensure that the communication between the electrodes and the scalp is established. For that purpose, an assessment of the impedance is made for each electrode. The lower it is, the better it is. In the case of wet electrodes, the experimenter has to inject gel at each sensor location. Once the impedances are lower than a certain threshold, typically a few kOhms, then the experiment can start. Regarding MEG, the experimenter places head-tracking coils to measure the head position before each recording. It helps preventing from large

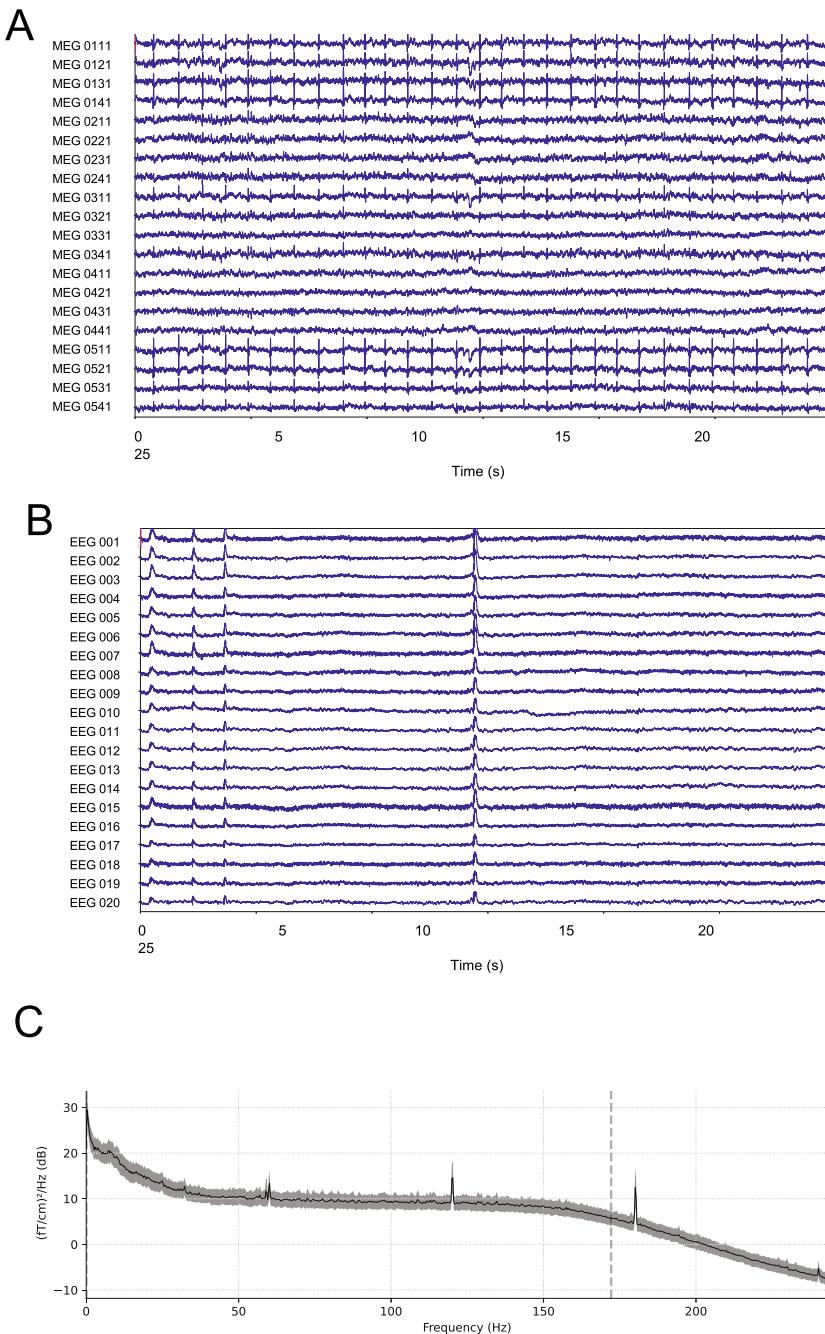


Fig. 5 Examples of artifacts in M/EEG. **(a)** Cardiac artifacts recorded with magnetometers. **(b)** Ocular artifacts recorded with EEG. **(c)** Power line noise recorded with gradiometers. Given its characteristics, plotting the power spectra enables to elicit it easily. The code to generate this figure is accessible via the dedicated [Github repository](#)

head movements that could lead to motion artefacts and error in the localization of source activity. The locations of fiducial points (nasion, left and right preauricular points) are registered. The information is stored in each data file. The subject is then placed in the magnetic shielding room after taking off all the elements that could generate magnetic interference with the device (e.g., jewels, belt). The experimenter helps the subject to place his/her head in the MEG helmet. Once the subject is in a comfortable position, the experimenter will save the head position that will be used as reference during the whole session.

Once the subject is correctly installed, the experimenter can start some pre-recordings to check the quality of the signal and give specific instructions to the subjects accordingly (e.g., loosening the jaw to avoid muscular artifacts). Finally, the experimenter can give further instructions regarding the task to perform before starting the recordings. After the end of session, the data are stored in specific servers to be processed.

4 Data Analysis

This section aims at providing recommendations for analyzing M/EEG data. An overview of the main steps of the M/EEG data analysis is provided in Fig. 6.

4.1 Types of Artifacts/Noise

The notion of artifacts depends strongly on the signal of interest. Here, we consider as artifacts the signals that make the recording more difficult and may hamper the analysis of the brain activity recorded with EEG and/or MEG. Such artifacts can be divided into two categories: the neurophysiological artifacts and the environmental noise. This section aims at presenting their main features.

4.1.1 Neurophysiological Artifacts

This category of artifacts corresponds to noise generated by the subjects themselves, whether it is voluntarily or not. In a nutshell, it is important to bear in mind that the brain is far from being the only organ that generates electromagnetic activity. In particular, the eyes and the heart produce electromagnetic activity, which shows an amplitude higher than that of the brain. As a result, the main neurophysiological artifacts are related to cardiac activity and ocular activity (via blinks and saccades) and can be visually spotted out during an M/EEG recording (*see* Fig. 5a, b). A possible way to reduce the ocular artifacts is to instruct the subject to avoid moving their eyes and, for short recordings only, to avoid eyes blinking. Another neurophysiological artifact may be induced by the subjects' voluntarily motion. Indeed, motion engenders muscular activity that can distort the recorded brain signals. Typical examples are jaw clenching and swallowing. They generate high-frequency

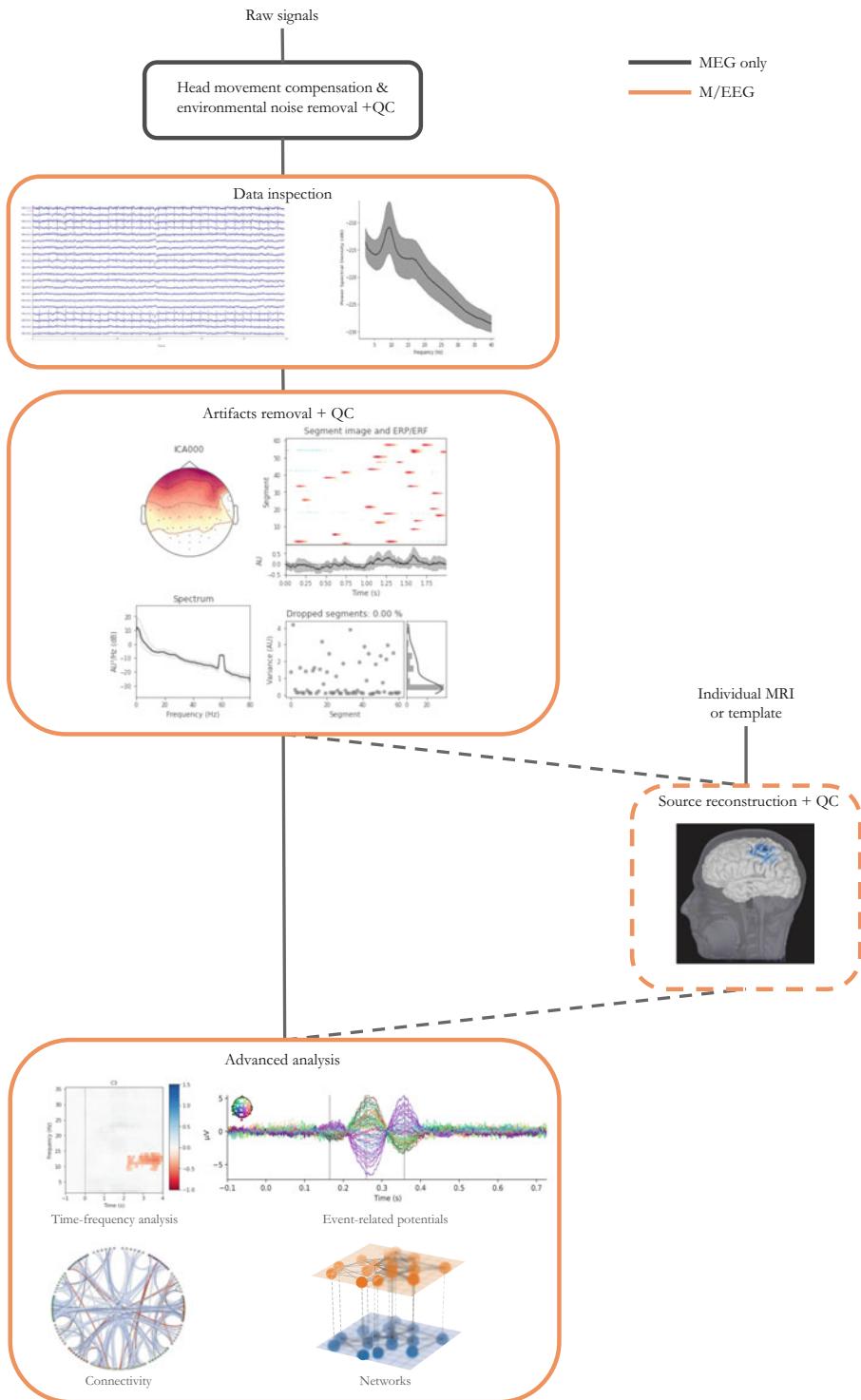


Fig. 6 Data analysis in M/EEG: general workflow. QC stands for quality check. Source reconstruction is not compulsory but advisable in specific cases. The code to generate this figure is accessible via the dedicated [Github repository](#)

activity that propagates to temporal electrodes. In the specific case of MEG, the device consisting in a helmet, it is strongly sensitive to head motion. A possible way to reduce the muscular artifacts is to instruct the subjects to remain as quiet as possible and to avoid moving their jaws.

4.1.2 Environmental Noise

This category refers to the artifacts generated by the environment that surrounds the experimental setup. They can be magnetic (e.g., magnetized devices that can interfere with the MEG sensors), linked with mechanical vibrations (e.g., presence of a tramway nearby), or simply associated with power line (occurring at 50 Hz or 60 Hz; *see* Fig. 5c). We do not aim at being exhaustive. We simply want the reader to be aware of the possible sources of environmental noise when analyzing M/EEG signals even though the Faraday cage and the shielded room, used respectively, in EEG and MEG, can partly prevent them.

4.1.3 System Noise

This category refers to artifacts generated by the sensors themselves. For example, in MEG, one can observe SQUID jumps or saturation. In both MEG and EEG, one can have broken sensors.

4.2 Preprocessing

This section aims at presenting the main steps that constitute the preprocessing pipeline, dedicated to artifact removal. This is probably the most crucial part when analyzing M/EEG data. Indeed, the point here is to remove noise without eliminating information of interest or distorting the signal. Attention must be paid to build the pipeline the most suited to the dataset and to the scientific question to be addressed. As such, the first thing to do when working with a new dataset is to extensively study it, in particular, inspecting the M/EEG signals but also the associated broadband power spectra. This preliminary step enables to identify most of the artifacts and, more importantly, if they have a specific temporal and/or frequency signature (e.g., presence of periodic artifacts).

From this point, it is possible to choose a specific strategy to remove the observed noise. In the case of cardiac and ocular artifacts, given their clear pattern, an efficient way to isolate and reduce them consists in applying independent component analysis (ICA) [13]. One can visually identify the components to be removed from both the temporal and the topographies (to avoid removing too many components) and manually select them. Another possibility, more reproducible, consists of using biosignals (e.g., electrocardiogram and electrooculogram) and to compute correlations between time series. This technique enables to ensure the robustness of the decision of removing a component.

In the case of artifacts at a specific frequency (e.g., power line noise at 50 Hz or 60 Hz), one can consider applying notch filters. With the same philosophy, in the case of muscular activity, applying a low-pass filter with a cutoff frequency at 40 Hz can be of interest.

Nevertheless, one objection can be raised: the signal distortion induced by the filtering. As previously explained, here, we aim at finding a trade-off between removing artifacts and preserving the information of interest. That is why the pipeline strongly depends on the scientific question to be addressed. In the case of muscular activity, if someone is interested in the activity in the gamma band (>30 Hz), applying a low-pass filter will be a poor choice, and as such, removing noisy trials can be an option. Regarding head motion, as explained in Subheading 3.2, MEG systems enable to register the head position. Methods relying notably on signal space separation [14] can correct small movements (i.e., less than several centimeters).

Another type of artifacts consists of a broken channel. To avoid having a different number of sensors from one subject to another, the proposed solution depends on the sensor location. If the sensor has four neighbors, strategies relying on the interpolation can be considered. It consists in creating a virtual sensor that is the linear combination of the signals recorded by the broken sensor's neighbors. If the sensor is located on the periphery, the interpolation is no longer reliable. The experimenter may consider removing the channel from the dataset. In the specific case of MEG, after an optional head movement correction step, if SQUID jump artifacts remain, one should consider reapplying the head movement correction on the raw data after having labeled as "bad" the sensors that show jumps. The bad MEG channels will be reconstructed.

Once the pipeline has been chosen and tested, it is important to check that the signals have been correctly preprocessed. This step corresponds to the quality check. There are different possibilities to perform it. The qualitative way would consist in superimposing preprocessed and postprocessed signals (which can be displayed as time series and/or power spectra) and to visualize potential differences. A more reliable way would consist in identifying a judgment criterion to assess to which extent the output signals are noisy. Possible metrics are the variance, the z-score, or the kurtosis. Using one of these metrics on the output may lead to both noisy channels and trials to be discarded. As a rule of thumb, the trial elimination must not exceed 10% of the total number of trials to ensure to have enough data to perform a relevant analysis [15].

4.3 Source Reconstruction

It is possible to directly analyze the signals recorded by the sensors. In such a case, one will say that the analysis is performed in the space of the sensors. However, it is also possible to go one step further and estimate the activity within the brain. This processing step is called source reconstruction and consists in estimating the neural correlates M/EEG signal location. It can be performed when one wants to have access to a higher spatial resolution to provide a more accurate description, and interpretation, of the neurophysiological phenomena occurring. For that purpose, both direct and inverse problems need to be solved [15, 16].

4.3.1 Direct Problem

Here, we aim at modeling the electromagnetic field produced by a cerebral source with known characteristics. For that purpose, it is necessary to consider both a physical model of the sources and a model that predicts the way that these sources will generate electromagnetic fields at the scalp level. The simplest model is the spherical model, which considers the head as an ensemble of spheres. Each sphere corresponds to a given tissue (brain, cerebro-spinal fluid, skull, or skin) characterized by a given conductivity. Even though it is possible to adjust the spheres to the geometry of the head or restrict them to a limited number of regions of interest, this model is an oversimplification of the head geometry. More realistic models rely on geometrical reconstruction of the different layers that form the head tissues, directly extracted from the anatomical magnetic resonance imaging (MRI) data of, ideally, the participant (the MRI thus needs to be acquired separately) or a dedicated template (e.g., MNI Colin 27). They consist in building meshes of the interfaces between different tissues. We can cite three approaches: the boundary element method (BEM) [17] that is the most widely used, the finite difference method (FDM), and the finite element method (FEM). Another model, called overlapping spheres [18], consists of fitting a given sphere under each sensor.

Even though there are no guidelines regarding the choice of the method, we could provide some elements of recommendations: given the high sensitivity of the EEG toward variations in terms of conductivity, the BEM model can be a tool of choice. As for the MEG, being less sensitive to changes in conductivity, the overlapping spheres can be considered.

4.3.2 Inverse Problem

One of the main challenges of the inverse problem lies in the nonuniqueness of its solution. In other words, a large number of brain activity patterns could generate the same signature detected at the sensor level. Therefore, some constraints or assumptions are essential to lead to a unique solution that reflects the best the acquired data [15, 16]. In this section, we aim at providing a short overview of the methods that are the most used in routine.

The dipole modeling methods rely on a source modeling via a reduced number of equivalent dipoles where each of them represents a source activity. As a result, such methods are based on an a priori hypothesis on the required number of sources.

Scanning methods, such as the MUSIC approach [19], consist in estimating the probability of presence of a current dipole inside each voxel. Among them are the beamformer methods [20], which consist in applying a spatial filtering to estimate the source activity at each location. We can cite the linearly constrained minimum variance (LCMV) and the synthetic aperture magnetometry (SAM) [21] as examples of beamformer methods [22].

The approaches relying on distributed source models consist in estimating the amplitudes of dipoles located on the cortical surface. The characteristics of the groups of dipoles are fixed or are estimated via the individual MRI of the participant. The most famous methods relying on distributed sources models are the weighted minimum norm (wMNE) [23, 24] and LORETA [25].

Similar to the preprocessing step, there is no ideal choice of method for the inverse problem, as it depends on the question to be addressed. A general recommendation would be to consider the minimum norm method when expecting distributed sources and the dipole modeling for focal sources.

5 Feature Extraction and Selection

When considering M/EEG from the machine learning perspective, an important aspect is the extraction and the selection of the features. This section aims at presenting the main features that can be extracted from M/EEG. As previously mentioned, the selection of the features depends on the scientific question to be addressed but also on the neurophysiological phenomenon underlying the M/EEG experiment. In M/EEG, filtering both in the time domain and in the spatial domain to select the most relevant features is common.

The two main types of features used in the literature rely on the information in the frequency domain and in the time domain. In an effort of completeness, we will see alternative features that reflect the interconnected nature of the brain.

The event-related features consist of chunks of time series concatenated from all the channels, resulting from a low-pass or band-pass filtering and/or from a down-sampling step. This category of features is relevant when considering evoked activity after the presentation of a given stimulus (e.g., visual, auditory, or sensory). They are therefore of interest when one is expecting significant changes in signal amplitudes occurring at a given moment. In the example presented in Subheading 2.2, a positive wave occurred 300 ms after the visual stimulation. One could consider using chunks of time series centered at $t = 300$ ms to detect automatically the P300 wave.

The spectral features are used in the case of the detection of an oscillatory activity (*see* Subheading 2.2), when changes in M/EEG rhythms amplitudes are expected. The features are associated with the power spectra estimated in a given channel and in a given frequency band for a specific time window. Power spectra can be computed via a plethora of methods; we can notably cite the spectrogram, the Morlet wavelet scalogram, and the auto-regressive models. For a thorough comparison of spectral feature extraction techniques on EEG signals, please refer to [26].

Spatial filtering can be a valuable tool both for event-related and spectral features [27]. It relies on the combination of signals, recorded from different sensors, to obtain a new one, associated with an improved signal-to-noise ratio. We can divide the spatial-filtering methods into three categories. The first one, not data-driven, relies on physical considerations regarding the way the signals propagate through the different brain tissues. The most famous illustration of this category is the Laplacian filter. In its simplest version, the small Laplacian consists, for each electrode location, of a derivation of the EEG waveform via the average signal computed from the four nearest neighbors [28]. The second category of spatial filtering is data-driven and unsupervised. It can rely, for example, on a principal component analysis (PCA) approach (*see* Chap. 2, Sect. 13.1). The third category is data-driven and supervised. The most famous examples in M/EEG are the common spatial patterns (CSP) for spectral features [29] and xDAWN for event-related features [30]. The CSP consists of a linear combination of EEG signals to maximize the difference between two classes in terms of variance. The xDAWN approach aims at improving the signal-to-noise ratio obtained with evoked potentials via a projection of the raw EEG signals onto an estimated evoked subspace. Recent efforts have been put together to combine approaches to provide ways to optimize simultaneously spectral and spatial filters, with, for example, the filter bank CSP (FBCSP) [31].

Even though spectral and event-related features are the most used in the M/EEG literature, alternative features have been considered in the past years. Firstly, features relying on covariance matrices have recently been extensively used, in particular for Riemannian geometry-based classification [32]. Despite an unclear neurophysiological interpretation, they enabled to reach state-of-the-art performance and to win a large number of competitions. Secondly, new features, which take into account the interconnected nature of brain functioning, have recently emerged [33]. There is a plethora of estimators to assess the intensity of the interactions between brain areas [34]. The most frequent estimators used as features in M/EEG are derived from the coherency, i.e., the normalized cross-spectral density obtained from two signals (e.g., imaginary part of coherence), or rely on the assessment of the phase synchrony between two signals (e.g., phase-locking value (PLV), phase-lag index (PLI)). Here, two challenges need to be dealt with: the volume conduction that can lead to spurious connectivity¹ and the online implementation. In the first case, even though some estimators, such as the imaginary coherence, are less sensitive to the volume conduction, working in the source space is recommended. In the second case, a large majority of studies that

¹ Originating from the mixture of signals engendered by different sources recorded at a given sensor.

consider estimators of functional interactions between two brain areas (i.e., functional connectivity estimators) as features are performed offline. Estimating brain interactions in real time is not trivial: it consists in finding the compromise between ensuring the quasi-stationarity of the signals and the statistical reliability of the functional connectivity estimation [33]. Recent studies considered the use of brain network metrics as potential features. Again, there is a plethora of metrics that characterize brain networks [33]. Here, we will cite the most used metrics. At the local scale, the node degree counts the number of connections linking one node to the other. In weighted networks (i.e., without having filtered the connectivity/adjacency matrix), it is referred to as node strength and consists in summing the weights of the connections of the considered node [35]. Another local-scale property of interest is the betweenness centrality defined as the extent to which a node lies “between” other pairs of nodes via the proportion of shortest paths in the network passing through it. This metric enables the identification of the nodes that are crucial for the information transfer between distant regions. At the global scale, we can cite two metrics: the characteristic path length and the clustering coefficient. The characteristic path length indicates the global tendency of the nodes in the network to integrate and exchange information. The clustering coefficient measures the tendency of having nodes’ neighbors mutually interconnected. Lastly, it is worthwhile noting the use of heterogeneous features (e.g., relying on both functional estimators and power spectra) that improves the classification accuracy [27]. Such an approach leads to an increase of the dimension, requiring cautions to select the most relevant features, via dimensional reduction methods.

The feature selection is a crucial step as it prevents redundancy, ensures the reliability of the features, reduces the dimensionality tuned, and helps in providing interpretable results. In this section, we aim at presenting the most popular feature selection methods in the M/EEG domain. For a complete description of the feature selection methods, the reader can refer to [27]. They can be divided into three categories: embedded, filter, and wrapper methods. In filter methods, the feature selection is performed independently and before the evaluation. Different criteria can be chosen to select features. The most popular criterion is the R^2 score, which assesses to which extent a given feature is influenced by a task performed by the subject. In wrapper methods, the feature selection utilizes the classification. In other words, in an iterative process, the relevance of each subset of features is assessed via the classification performance until a given criterion is met. The embedded method consists in integrating both the feature selection and the classification in the same process, via a decision tree, for example, or an ℓ_1 penalty term.

Box 1: Tools for M/EEG analysis

All these tools provide a wide range of tutorials, publicly available datasets, and codes.

Python-based:

- MNE-Python [36]
- MOABB [37]

MATLAB-based:

- EEGLAB [38]
- Fieldtrip [39]
- Brainstorm [40]
- SPM [41]

6 M/EEG and Brain Disorders

6.1 Clinical Applications of M/EEG

The spatial and temporal resolutions of M/EEG enable the observation of a large number of processes. Notably, they can detect both evoked responses and oscillatory activity. As such, using these information could pave the way to biomarkers of brain disorders. To illustrate this point, we will focus our presentation on two specific clinical applications: epilepsy and Alzheimer disease. Nevertheless, M/EEG can be useful for a wider range of applications both in neurological and psychiatric disorders [42, 43].

6.1.1 Epilepsy

Epilepsy is a neurological disorder that presents a high prevalence of 1% [44]. It is established that between 20 and 30% of the patients present a pharmacoresistant form of epilepsy [45]. Among this proportion of patients, only 30% can undergo a surgery [46]. Epilepsy is a distributed disease that induces brain network reorganization and brain rhythm alterations both during ictal and interictal periods [47, 48]. Due to its time resolution compatible with the capture of dynamical changes as well as its wide availability, EEG is a key modality for the evaluation of epilepsy [44]. In addition to scalp EEG, stereotactic-EEG (SEEG) can be used to further localize epileptogenic foci and proven to provide valuable information on epileptogenic networks [48]. MEG can also be used for pre-surgical evaluation and for functional mapping [49], but it is much more costly and less widely available.

The use of network theory in epilepsy provides a useful framework to characterize the seizure (onset and propagation), and its clinical expression (e.g., comorbidities) [47, 48]. At the local scale,

the node strength or degrees and the betweenness centrality have been used to characterize the epileptic network [48, 50]. At the global scale, two metrics have proven to be of interest in epilepsy [51]: the characteristic path length and the clustering coefficient.

6.1.2 Alzheimer Disease

Alzheimer disease is the most common dementia with 60–80% of the cases. The first symptoms are a deficit in short-term memory and concentration, followed later by a decline of linguistic skills, visuospatial orientation, and abstract reasoning judgment. As the pathophysiological process of the disease starts many years before the occurrence of symptoms [52], it is crucial to elicit biomarkers to provide a diagnosis as soon as possible. Efforts have been put together to describe mild cognitive impairment (MCI) and Alzheimer disease (AD) with M/EEG. These studies are essentially focused on oscillatory activity and on interactions between brain areas [53, 54]. In particular, patients present a reduced synchrony [55], and a decrease of the alpha power (i.e., between 8 and 12 Hz) correlates with lower cognitive status and hippocampal atrophy. Studies performed with MEG in preclinical and prodromal stages of AD showed that the effects of amyloid-beta deposition were associated with an increment of the prefrontal alpha power and that altered connectivity in the default mode network was present in normal individuals at risk for AD [56, 57].

A recent EEG work showed that effects of neurodegeneration were focused in frontocentral regions with an increase in high-frequency bands (beta and gamma) and a decrease in lower-frequency bands (delta) [58]. In particular, EEG patterns differ depending on the degree of amyloid burden, suggesting a compensatory mechanism: following a U-shape curve in delta power and an inverted U-shape curve for other tested metrics.

6.2 Advanced Uses: The Example of BCI as a Rehabilitation Tool

6.2.1 Presentation of the BCI

Brain-computer interfaces (BCIs) consist of acquiring, analyzing, and translating brain signals into commands in real time for control or communication. These systems present a large number of clinical applications and assistive technologies including control of wheelchairs and brain-based communication. BCI devices can be a valuable tool in the treatment of neurological disorders such as stroke [59] and to provide assistive solutions for patients with spinal cord injury [60] or the amyotrophic lateral sclerosis [59]. With regard to the communication, devices such as the P300 Speller, which rely on the evoked response occurring 300 ms after the visual stimulation, allow the users to communicate by selecting letters to form words and even sentences. For an overview of the main steps to be considered when performing a BCI experiment, please refer to Fig. 7.

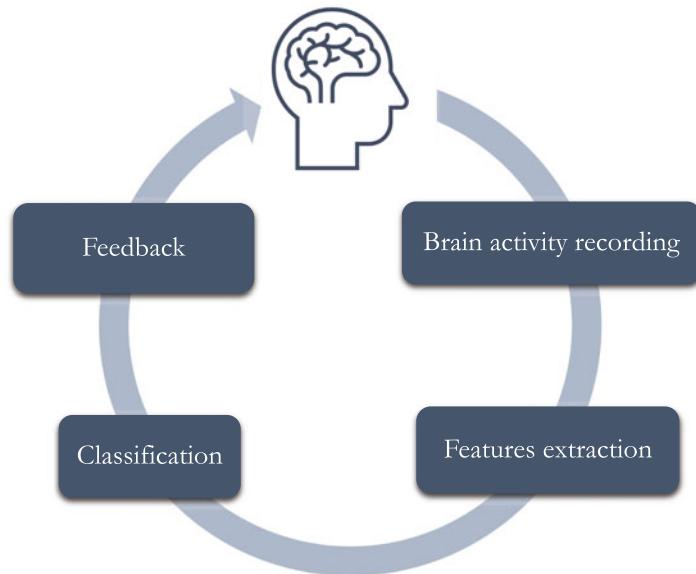


Fig. 7 BCI experiment workflow

6.2.2 BCI as a Rehabilitation Tool

Stroke is one of the most common neurological conditions. In 2010, stroke was the second leading cause of death worldwide [61]. After a stroke, most patients require rehabilitation and assistance for daily tasks. Motor deficit of the upper limbs affects 70% of the survivors [62], and 85% of those presenting paralysis will have persistent damage [63]. Rapid recovery is observed during the first 3 months (*acute phase*) but can continue for several months after the accident (*chronic phase*) [64]. Motor imagery (MI)-based BCI can constitute a motor substitution in the case of stroke by building alternative pathways from the stimulation to the brain [65]. In this particular case, the system relies on the desynchronization effect associated with a decrease of the power spectra computed within the contralateral sensorimotor area [66]. In a recent meta-analysis [67], the authors observed that rehabilitation to restore upper limb motor function based on BCIs could improve the motricity, assessed via the Fugl-Meyer scale, more than other therapies. A part of the screened studies showed that BCI could induce neuroplasticity.

Brain network changes in stroke patients represent a very promising clinical application of closed-loop systems in rehabilitation strategies. Motor imagery has been proven to be a valuable tool in the study of upper limb recovery after stroke [68]. It enabled observations of changes in ipsilesional intrahemispheric connectivity [69] but also modifications in connectivity in prefrontal areas and correlations between node strengths and motor outcome [70]. Based on previous observations in resting state [71], a recent double-blind study involving ten stroke patients at the chronic stage revealed that node strength, computed from the

ipsilesional primary motor cortex in the alpha band, could be a target for a motor imagery-based neurofeedback and lead to significant improvement on motor performance [72].

6.2.3 Current Challenges and Perspectives

Despite being beneficial for patients, controlling a BCI system is a learned skill that 15–30% of the users cannot develop even after several training sessions. This phenomenon, called “BCI inefficiency” [73], has been presented as one of the main limitations to a wider use of BCI. From the machine learning perspective, the main challenges to overcome in current BCI paradigms relying on EEG recordings are the low signal-to-noise ratio of signals, the non-stationarity over time mainly resulting from the difference between calibration and feedback sessions, the reduced amount of available data to train the classifier explained by the number of classes to be discriminated and/or the need to avoid the subject’s tiredness, and the lack of robustness and reliability of the BCI systems, in particular when decoding the users’ mental command.

To tackle these challenges, efforts have been put to improve the classification algorithms. They can be divided into three main groups: the adaptive classifiers, the transfer learning techniques, and the matrix- or tensor-based algorithms. The adaptive classifiers aim at dealing with EEG non-stationarity by taking into account changes in signal properties, and feature distribution, over time. Their parameters are updated when new EEG signals are available [74]. Even though most of the adaptive classifiers can rely on a supervised approach, the unsupervised one has proven to outperform the classifiers that cannot catch temporal dynamics [75]. Besides, it can be a valuable tool to reduce the training duration and potentially to remove the calibration part. Nevertheless, the adaptive classifiers present one main pitfall: their lack of online validation with a user in most of the current literature. This leads to two potential issues: the difficulty to find a trade-off between fully retraining the classifier and updating some key parameters and the adaptation that may not follow the actual user’s intent by being too fast or too slow [76].

Transfer learning consists here in exploiting changes in EEG signal properties over time and subjects to extract knowledge. More specifically, it relies on learned classifiers that are trained on one task (called domain here) and are adapted to another task with little or no new training data [77]. For example, it can be applied to a dataset formed by two motor imagery tasks performed by two different subjects. There is plethora of methods to solve the transfer learning problem [78]. The most common in the EEG-based BCI domain consists in learning the transformation to correct the mismatch between the domains, occurring when one domain corresponds to a hand motor imagery and the other to a foot motor imagery, for instance, finding a common feature representation for the domains, or learning a transformation of the data to make their

distribution match [27]. Despite its robustness and recent advances in proposing guidelines [79], there is a lack of online experiments relying on transfer learning to fully validate this approach and assess to which extent it can be beneficial to patients.

Among the classification methods relying on matrices and tensors, the most well-known is the Riemannian geometry-based one. One of the main original characteristics of this approach is that it is able to manipulate and classify the data by representing them as symmetric positive definite matrices, such as covariance matrices, and by mapping them onto a dedicated geometrical space, involving less steps than the classic approaches. This approach relies on the assumption that the sources are specific of a given task encoded via the covariance matrix computed from EEG signals. Here, trials are classified via nearest neighbor methods relying on the Riemannian distance and the geometric mean. With the method relying on the minimum distance to mean (MDM), each class is associated with a geometric mean computed from the training data. Then, the MDM will attribute an unlabeled trial to the class showing the closest mean [80]. The Riemannian approaches present many advantages: they can be applied to all BCI paradigms, no parameter tuning is required, they are robust to noise, and, combined to transfer learning methods, they can lead to calibration-free BCI sessions [81]. In particular, Riemannian geometry-based methods [80, 82] are now the state of the art in terms of performance [27] and have won several data competitions² [83].

Box: 2 To go further

Guidelines and books of reference

- Hari, M., and Puce, A. (2017). MEG-EEG Primer. In MEG-EEG Primer. Oxford University Press.
- M. Clerc, L. Bougrain, and F. Lotte. (2016) Brain-Computer Interfaces 1: Methods and Perspectives, Wiley.
- M. Clerc, L. Bougrain, and F. Lotte. (2016) Brain-Computer Interfaces 2: Technology and Applications, Wiley.
- Gross, J., Baillet, S., Barnes, G. R., Henson, R. N., Hillebrand, A., Jensen, O., Jerbi, K., Litvak, V., Maess, B., Oostenveld, R., Parkkonen, L., Taylor, J. R., van Wassenhove, V., Wibral, M., and Schoffelen, J.-M. (2013). Good practice for conducting and reporting MEG research. *Neuroimage*, 65, 349–363.

(continued)

² See, for example, the 6 competitions won by A. Barachant: <http://alexandre.barachant.org/challenges/>.

Box 2 (continued)

- Puce, A. and Hämäläinen, M. S. (2017). A Review of Issues Related to Data Acquisition and Analysis in EEG/MEG Studies. *Brain Sci*, 7(6).

7 Conclusion

EEG and MEG are key modalities for the study of brain disorders. In particular, EEG is relatively cheap and widely available and is thus a widely used tool in neurology. When dealing with EEG and MEG data, it is important to understand the origin of the signals as well as the different steps in their preprocessing and feature extraction. Machine learning is increasingly used on EEG and MEG data, in particular for BCI but also for computer-aided diagnosis and prognosis of brain disorders.

Acknowledgements

The author acknowledges support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 864729) and from the program "Investissements d'avenir" ANR-10-IAIHU-06. The author thanks D. Schwartz for his suggestions and insightful comments.

References

- Seeck M, Koessler L, Bast T, Leijten F, Michel C, Baumgartner C, He B, Beniczky S (2017) The standardized EEG electrode array of the IFCN. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 128(10): 2070–2077. <https://doi.org/10.1016/j.clinph.2017.06.254>
- Casson AJ (2019) Wearable EEG and beyond. *Biomed Eng Lett* 9(1):53–71. <https://doi.org/10.1007/s13534-018-00093-6>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6431319/>
- Hämäläinen M, Hari R, Ilmoniemi RJ, Knuutila J, Lounasmaa OV (1993) Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev Mod Phys* 65(2):413–497. <https://doi.org/10.1103/RevModPhys.65.413> <https://link.aps.org/doi/10.1103/RevModPhys.65.413>
- Cohen D (1972) Magnetocephalography: detection of the brain's electrical activity with a superconducting magnetometer. *Science* 175(4022):664–666
- Vrba J, Robinson SE (2001) Signal Processing in Magnetoencephalography. *Methods* 25(2): 249–271. <https://doi.org/10.1006/meth.2001.1238>. <https://www.sciencedirect.com/science/article/pii/S1046202301912381>
- Corsi MC (2015) Magnétomètres à pompage optique à Hélium 4: développement et preuve de concept en magnétocardiographie et en magnétoencéphalographie. PhD thesis, Grenoble Alpes. <http://www.theses.fr/2015GREAT082>

7. Tierney TM, Holmes N, Mellor S, López JD, Roberts G, Hill RM, Boto E, Leggett J, Shah V, Brookes MJ, Bowtell R, Barnes GR (2019) Optically pumped magnetometers: From quantum origins to multi-channel magnetoencephalography. *NeuroImage* 199: 598–608. <https://doi.org/10.1016/j.neuroimage.2019.05.063>. <http://www.sciencedirect.com/science/article/pii/S1053811919304550>
8. Boto E, Seedat ZA, Holmes N, Leggett J, Hill RM, Roberts G, Shah V, Fromhold TM, Mullinger KJ, Tierney TM, Barnes GR, Bowtell R, Brookes MJ (2019) Wearable neuroimaging: combining and contrasting magnetoencephalography and electroencephalography. *NeuroImage* 201:116099. <https://doi.org/10.1016/j.neuroimage.2019.116099>. <http://www.sciencedirect.com/science/article/pii/S1053811919306901>
9. Barry DN, Tierney TM, Holmes N, Boto E, Roberts G, Leggett J, Bowtell R, Brookes MJ, Barnes GR, Maguire EA (2019) Imaging the human hippocampus with optically-pumped magnetoencephalography. *NeuroImage* 203: 116192. <https://doi.org/10.1016/j.neuroimage.2019.116192>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6854457/>
10. Boto E, Holmes N, Leggett J, Roberts G, Shah V, Meyer SS, Muñoz LD, Mullinger KJ, Tierney TM, Bestmann S, Barnes GR, Bowtell R, Brookes MJ (2018) Moving magnetoencephalography towards real-world applications with a wearable system. *Nature* 555(7698):657–661. <https://doi.org/10.1038/nature26147>. <https://www.nature.com/articles/nature26147>
11. Brookes MJ, Boto E, Rea M, Shah V, Osborne J, Holmes N, Hill RM, Leggett J, Rhodes N, Bowtell R (2021) Theoretical advantages of a triaxial optically pumped magnetometer magnetoencephalography system. *NeuroImage* 236:118025. <https://doi.org/10.1016/j.neuroimage.2021.118025>. <https://www.sciencedirect.com/science/article/pii/S1053811921003025>
12. Labyt E, Corsi MC, Fourcault W, Laloy AP, Bertrand F, Lenouvel F, Cauffet G, Prado ML, Berger F, Morales S (2019) Magnetoencephalography with optically pumped 4He magnetometers at Ambient temperature. *IEEE Trans Med Imaging* 38(1):90–98. <https://doi.org/10.1109/TMI.2018.2856367>. Conference name: IEEE Trans Med Imaging
13. Bell AJ, Sejnowski TJ (1995) An information-maximization approach to blind separation and blind deconvolution. *Neural Comput* 7(6): 1129–1159. <https://doi.org/10.1162/neco.1995.7.6.1129>
14. Taulu S, Simola J (2006) Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Phys Med Biol* 51(7):1759–1768. <https://doi.org/10.1088/0031-9155/51/7/008>
15. Gross J, Baillet S, Barnes GR, Henson RN, Hillebrand A, Jensen O, Jerbi K, Litvak V, Maess B, Oostenveld R, Parkkonen L, Taylor JR, van Wassenhove V, Wibral M, Schoffelen JM (2013) Good practice for conducting and reporting MEG research. *NeuroImage* 65: 349–363. <https://doi.org/10.1016/j.neuroimage.2012.10.001>
16. Baillet S, Mosher J, Leahy R (2001) Electromagnetic brain mapping. *IEEE Signal Process Mag* 18(6):14–30. <https://doi.org/10.1109/79.962275>
17. Fuchs M, Wagner M, Kastner J (2001) Boundary element method volume conductor models for EEG source reconstruction. *Clin Neurophysiol* 112(8):1400–1407. [https://doi.org/10.1016/S1388-2457\(01\)00589-2](https://doi.org/10.1016/S1388-2457(01)00589-2). <http://www.sciencedirect.com/science/article/pii/S1388245701005892>
18. Huang MX, Mosher JC, Leahy RM (1999) A sensor-weighted overlapping-sphere head model and exhaustive head model comparison for MEG. *Phys Med Biol* 44(2):423–440. <https://doi.org/10.1088/0031-9155/44/2/010>
19. Mosher J, Baillet S, Leahy R (1999) EEG source localization and imaging using multiple signal classification approaches. *J Clin Neurophysiol* 16(3):225–238. <https://doi.org/10.1097/00004691-199905000-00004>
20. Hillebrand A, Barnes GR (2005) Beamformer Analysis of MEG Data. In: International review of neurobiology, magnetoencephalography, vol 68. Academic Press, New York, pp 149–171. [https://doi.org/10.1016/S0074-7742\(05\)68006-3](https://doi.org/10.1016/S0074-7742(05)68006-3). <https://www.sciencedirect.com/science/article/pii/S0074774205680063>
21. Robinson S, Vrba J (1999) Functional neuro-imaging by synthetic aperture magnetometry (SAM)—ScienceOpen. Recent advances in Biomagnetism. Tohoku University Press, Sendai, pp 302–305. <https://www.scienceopen.com/document?vid=067e8134-c846-4b5f-9ae1-656869ca8727>
22. Van Veen B, Van Drongelen W, Yuchtman M, Suzuki A (1997) Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Trans Biomed Eng* 44(9):867–880. <https://doi.org/10.1109/10.623056>

23. Fuchs M, Wagner M, Köhler T, Wischmann HA (1999) Linear and nonlinear current density reconstructions. *Journal of Clinical Neurophysiology: Official Publication of the American Electroencephalographic Society* 16(3):267–295
24. Lin FH, Witzel T, Ahlfors SP, Stufflebeam SM, Belliveau JW, Hämäläinen MS (2006) Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. *NeuroImage* 31(1):160–171. <https://doi.org/10.1016/j.neuroimage.2005.11.054>
25. Pascual-Marqui RD (2002) Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find Exp Clin Pharmacol* 24(Suppl D):5–12
26. Brodu N, Lotte F, Lécuyer A (2011) Comparative study of band-power extraction techniques for Motor Imagery classification. In: 2011 IEEE symposium on computational intelligence, cognitive algorithms, mind, and brain (CCMB), pp 1–6. <https://doi.org/10.1109/CCMB.2011.5952105>
27. Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A, Yger F (2018) A review of classification algorithms for EEG-based brain-computer interfaces: a 10-year update. *J Neural Eng* 15(3):031005. <https://doi.org/10.1088/1741-2552/aab2f2>
28. McFarland DJ, McCane LM, David SV, Wolpaw JR (1997) Spatial filter selection for EEG-based communication. *Electroencephalogr Clin Neurophysiol* 103(3):386–394. [https://doi.org/10.1016/S0013-4694\(97\)00222-2](https://doi.org/10.1016/S0013-4694(97)00222-2). <http://www.sciencedirect.com/science/article/pii/S001346949700222>
29. Ramoser H, Muller-Gerking J, Pfurtscheller G (2000) Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans Rehabil Eng* 8(4):441–446. <https://doi.org/10.1109/86.895946>
30. Rivet* B, Souloumiac A, Attina V, Gibert G (2009) xDAWN Algorithm to enhance evoked potentials: application to brain-computer interface. *IEEE Trans Biomed Eng* 56(8): 2035–2043. <https://doi.org/10.1109/TBME.2009.2012869>. Conference name: IEEE Trans Biomed Eng
31. Ang KK, Chin ZY, Zhang H, Guan C (2008) Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), pp 2390–2397. <https://doi.org/10.1109/IJCNN.2008.4634130>. iSSN: 2161-4407
32. Yger F, Berar M, Lotte F (2017) Riemannian approaches in brain-computer interfaces: a review. *IEEE Trans Neural Syst Rehabil Eng* 25(10):1753–1762. <https://doi.org/10.1109/TNSRE.2016.2627016>
33. Gonzalez-Astudillo J, Cattai T, Bassignana G, Corsi MC, De Vico Fallani F (2020) Network-based brain computer interfaces: principles and applications. *J Neural Eng* 18(1):011001. <https://doi.org/10.1088/1741-2552/abc760>. <http://iopscience.iop.org/article/10.1088/1741-2552/abc760>
34. Bastos AM, Schoffelen JM (2016) A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front Syst Neurosci* 9:175. <https://doi.org/10.3389/fnsys.2015.00175>. <https://www.frontiersin.org/articles/10.3389/fnsys.2015.00175/full>
35. Fornito A, Zalesky A, Bullmore E (2016) Fundamentals of brain network analysis, reprint edizione edn. Academic Press, Amsterdam
36. Gramfort A, Luesi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hämäläinen MS (2014) MNE software for processing MEG and EEG data. *NeuroImage* 86:446–460. <https://doi.org/10.1016/j.neuroimage.2013.10.027>
37. Jayaram V, Barachant A (2018) MOABB: trustworthy algorithm benchmarking for BCIs. *J Neural Eng* 15(6):066011. <https://doi.org/10.1088/1741-2552/aadea0>. Publisher: IOP Publishing
38. Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* 134(1): 9–21. <https://doi.org/10.1016/j.jneumeth.2003.10.009>
39. Oostenveld R, Fries P, Maris E, Schoffelen JM (2010) FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* 2011:e156869. <https://doi.org/10.1155/2011/156869>
40. Tadel F, Baillet S, Mosher J, Pantazis D, Leahy R (2011) Brainstorm: a user-friendly application for MEG/EEG analysis. *Comput Intell Neurosci* 2011:1–13. <https://doi.org/10.1155/2011/879716>
41. Penny WD, Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE (eds) (2006) Statistical parametric mapping: the analysis of functional brain images, illustrated edn. Academic Press, Amsterdam
42. Hari R, Baillet S, Barnes G, Burgess R, Forss N, Gross J, Hämäläinen M, Jensen O, Kakigi R,

- Mauguière F, Nakasato N, Puce A, Romani GL, Schnitzler A, Taulu S (2018) IFCN-endorsed practical guidelines for clinical magnetoencephalography (MEG). *Clin Neurophysiol* 129(8):1720–1747. <https://doi.org/10.1016/j.clinph.2018.03.042>. <http://www.sciencedirect.com/science/article/pii/S1388245718306576>
43. Boutros NN, Galderisi S, Pogarell O, Riggio S (2011) Standard electroencephalography in clinical psychiatry: a practical handbook. Wiley, London
44. des Enseignants de Neurologie C (2016) Épilepsies de l'enfant et de l'adulte. <https://www.cen-neurologie.fr/deuxieme-cycle/epilepsies-enfant-ladulte>
45. Kwan P, Brodie MJ (2000) Early identification of refractory epilepsy. *N Engl J Med* 342(5): 314–319. <https://doi.org/10.1056/NEJM200002033420503>
46. Micoulaud-Franchi JA, Lanteaume L, Pallanca O, Vion-Dury J, Bartolomei F (2014) Biofeedback et épilepsie pharmacorésistante: le retour d'une thérapeutique ancienne? *Rev Neurol* 170(3):187–196. <https://doi.org/10.1016/j.neurol.2013.10.011>
47. Diessen Ev, Diederend SJH, Braun KPJ, Jansen FE, Stam CJ (2013) Functional and structural brain networks in epilepsy: What have we learned? *Epilepsia* 54(11):1855–1865. <https://doi.org/10.1111/epi.12350>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/epi.12350>
48. Bartolomei F, Lagarde S, Wendling F, McGonigal A, Jirsa V, Guye M, Bénar C (2017) Defining epileptogenic networks: contribution of SEEG and signal analysis. *Epilepsia* 58(7):1131–1147. <https://doi.org/10.1111/epi.13791>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/epi.13791>
49. Stefan H, Trinka E (2017) Magnetoencephalography (MEG): past, current and future perspectives for improved differentiation and treatment of epilepsies. *Seizure* 44:121–124. <https://doi.org/10.1016/j.seizure.2016.10.028>. <http://www.sciencedirect.com/science/article/pii/S1059131116302217>
50. Wilke C, Worrell G, He B (2011) Graph analysis of epileptogenic networks in human partial epilepsy. *Epilepsia* 52(1):84–93. <https://doi.org/10.1111/j.1528-1167.2010.02785.x>
51. Bartolomei F, Bettus G, Stam CJ, Guye M (2013) Interictal network properties in mesial temporal lobe epilepsy: a graph theoretical study from intracerebral recordings. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 124(12):2345–2353. <https://doi.org/10.1016/j.clinph.2013.06.003>
52. Bateman RJ, Xiong C, Benzinger TLS, Fagan AM, Goate A, Fox NC, Marcus DS, Cairns NJ, Xie X, Blazey TM, Holtzman DM, Santacruz A, Buckles V, Oliver A, Moulder K, Aisen PS, Ghetti B, Klunk WE, McDade E, Martins RN, Masters CL, Mayeux R, Ringman JM, Rossor MN, Schofield PR, Sperling RA, Salloway S, Morris JC, Dominantly Inherited Alzheimer Network (2012) Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *N Engl J Med* 367(9): 795–804. <https://doi.org/10.1056/NEJMoa1202753>
53. Babiloni C, Lizio R, Marzano N, Capotosto P, Soricelli A, Triggiani AI, Cordone S, Gesualdo L, Del Percio C (2016) Brain neural synchronization and functional coupling in Alzheimer's disease as revealed by resting state EEG rhythms. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology* 103: 88–102. <https://doi.org/10.1016/j.ijpsycho.2015.02.008>
54. Engels MMA, van der Flier WM, Stam CJ, Hillebrand A, Scheltens P, van Straaten ECW (2017) Alzheimer's disease: the state of the art in resting-state magnetoencephalography. *Clinical Neurophysiology* 128(8):1426–1437. <https://doi.org/10.1016/j.clinph.2017.05.012>. <https://www.sciencedirect.com/science/article/pii/S1388245717301979>
55. Stam CJ (2010) Use of magnetoencephalography (MEG) to study functional brain networks in neurodegenerative disorders. *J Neurol Sci* 289(1):128–134. <https://doi.org/10.1016/j.jns.2009.08.028>. [https://www.jns-journal.com/article/S0022-510X\(09\)00784-9/abstract](https://www.jns-journal.com/article/S0022-510X(09)00784-9/abstract)
56. Nakamura A, Cuesta P, Kato T, Arahata Y, Iwata K, Yamagishi M, Kuratsubo I, Kato K, Bundo M, Diers K, Fernández A, Maestú F, Ito K (2017) Early functional network alterations in asymptomatic elders at risk for Alzheimer's disease. *Sci Rep* 7(1):6517. <https://doi.org/10.1038/s41598-017-06876-8>
57. Nakamura A, Cuesta P, Fernández A, Arahata Y, Iwata K, Kuratsubo I, Bundo M, Hattori H, Sakurai T, Fukuda K, Washimi Y, Endo H, Takeda A, Diers K, Bajo R, Maestú F, Ito K, Kato T (2018) Electromagnetic signatures of the preclinical and prodromal stages of Alzheimer's disease. *Brain J Neurol* 141(5): 1470–1485. <https://doi.org/10.1093/brain/awy044>

58. Gaubert S, Raimondo F, Houot M, Corsi MC, Naccache L, Diego Sitt J, Hermann B, Oudiette D, Gagliardi G, Habert MO, Dubois B, De Vico Fallani F, Bakardjian H, Epelbaum S (2019) EEG evidence of compensatory mechanisms in preclinical Alzheimer's disease. *Brain* 142(7):2096–2112. <https://doi.org/10.1093/brain/awz150>. <https://academic.oup.com/brain/article/142/7/2096/551996>
59. Kashihara K (2014) A brain-computer interface for potential non-verbal facial communication based on EEG signals related to specific emotions. *Front Neurosci* 8:244. <https://doi.org/10.3389/fnins.2014.00244>
60. King CE, Wang PT, Chui LA, Do AH, Nenadic Z (2013) Operation of a brain-computer interface walking simulator for individuals with spinal cord injury. *J Neuroeng Rehabil* 10:77. <https://doi.org/10.1186/1743-0003-10-77>
61. Feigin VL, Forouzanfar MH, Krishnamurthi R, Mensah GA, Connor M, Bennett DA, Moran AE, Sacco RL, Anderson L, Truelsen T, O'Donnell M, Venketasubramanian N, Barker-Collo S, Lawes CMM, Wang W, Shinozawa Y, Witt E, Ezzati M, Naghavi M, Murray C (2014) Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010. *Lancet* 383(9913):245–254. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4181600/>
62. Houwink A, Roorda LD, Smits W, Molenaar IW, Geurts AC (2011) Measuring upper limb capacity in patients after stroke: reliability and validity of the stroke upper limb capacity scale. *Arch Phys Med Rehabil* 92(9):1418–1422. <https://doi.org/10.1016/j.apmr.2011.03.028>. <https://linkinghub.elsevier.com/retrieve/pii/S0003999311002218>
63. Kavanagh S, Knapp M, Patel A (1999) Costs and disability among stroke patients. *J Public Health* 21(4):385–394. <https://doi.org/10.1093/pubmed/21.4.385>. <https://academic.oup.com/jpubhealth/article-lookup/doi/10.1093/pubmed/21.4.385>
64. de Santé HA (2012) Accident vasculaire cérébral: méthodes de rééducation de la fonction motrice chez l'adulte. Tech. rep. https://www.has-sante.fr/jcms/c_1334330/fr/accident-vasculaire-cerebral-methodes-de-reeducation-de-la-fonction-motrice-chez-l-adulte
65. Prasad G, Herman P, Coyle D, McDonough S, Crosbie J (2010) Applying a brain-computer interface to support motor imagery practice in people with stroke for upper limb recovery: a feasibility study. *J Neuroeng Rehabil* 7:60. <https://doi.org/10.1186/1743-0003-7-60>
66. Pfurtscheller G, Lopes da Silva FH (1999) Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin Neurophysiol* 110(11):1842–1857. [https://doi.org/10.1016/S1388-2457\(99\)00141-8](https://doi.org/10.1016/S1388-2457(99)00141-8). <http://www.sciencedirect.com/science/article/pii/S1388245799001418>
67. Cervera MA, Soekadar SR, Ushiba J, Millán JdR, Liu M, Birbaumer N, Garipelli G (2018) Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis. *Ann Clin Transl Neurol* 5(5):651–663. <https://doi.org/10.1002/acn3.544>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5945970/>
68. Sharma N, Pomeroy VM, Baron JC (2006) Motor imagery: a backdoor to the motor system after stroke? *Stroke* 37(7):1941–1952. <https://doi.org/10.1161/01.STR.0000226902.43357.fc>
69. Pichiorri F, Morone G, Petti M, Toppi J, Pisotta I, Molinari M, Paolucci S, Inghilleri M, Astolfi L, Cincotti F, Mattia D (2015) Brain-computer interface boosts motor imagery practice during stroke recovery. *Ann Neurol* 77(5):851–865. <https://doi.org/10.1002/ana.24390>. <https://www.readcube.com/articles/10.1002/ana.24390>
70. Sharma N, Baron JC, Rowe JB (2009) Motor imagery after stroke: relating outcome to motor network connectivity. *Ann Neurol* 66(5):604–616. <https://doi.org/10.1002/ana.21810>. <https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.21810>
71. Dubovik S, Pignat JM, Ptak R, Aboulafia T, Allet L, Gillabert N, Magnin C, Albert F, Momjian-Mayor I, Nahum L, Lascano AM, Michel CM, Schnider A, Guggisberg AG (2012) The behavioral significance of coherent resting-state oscillations after stroke. *NeuroImage* 61(1):249–257. <https://doi.org/10.1016/j.neuroimage.2012.03.024>
72. Mottaz A, Corbet T, Dogancı N, Magnin C, Nicolo P, Schnider A, Guggisberg AG (2018) Modulating functional connectivity after stroke with neurofeedback: effect on motor deficits in a controlled cross-over study. *NeuroImage: Clinical* 20:336–346. <https://doi.org/10.1016/j.nicl.2018.07.029>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6091229/>
73. Allison BZ, Neuper C (2010) Could Anyone Use a BCI? In: Tan DS, Nijholt A (eds) Brain-computer interfaces, human-computer interaction series. Springer London, pp 35–54. http://link.springer.com/chapter/10.1007/978-1-84996-272-8_3
74. Schlögl A, Vidaurre C, Müller KR (2010) Adaptive methods in BCI research—an

- introductory tutorial. In: Graimann B, Pfurtscheller G, Allison B (eds) Brain-computer interfaces: revolutionizing human-computer interaction, The Frontiers Collection. Springer, Berlin, pp 331–355. https://doi.org/10.1007/978-3-642-02091-9_18
75. Vidaurre C, Kawanabe M, von Bünau P, Blankertz B, Müller KR (2011) Toward unsupervised adaptation of LDA for brain-computer interfaces. IEEE Trans Biomed Eng 58(3):587–597. <https://doi.org/10.1109/TBME.2010.2093133>
76. Müller JS, Vidaurre C, Schreuder M, Meinecke FC, von Bünau P, Müller KR (2017) A mathematical model for the two-learners problem. J Neural Eng 14(3):036005. <https://doi.org/10.1088/1741-2552/aa620b>
77. Waytowich NR, Lawhern VJ, Bohannon AW, Ball KR, Lance BJ (2016) Spectral transfer learning using information geometry for a user-independent brain-computer interface. Front Neurosci 10:430. <https://doi.org/10.3389/fnins.2016.00430>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5032911/>
78. Pan SJ, Yang Q (2010) A Survey on Transfer Learning. IEEE Trans Knowl Data Eng 22(10):1345–1359. <https://doi.org/10.1109/TKDE.2009.191>
79. Bougrain L, Rimbert S, Rodrigues PLC, Canron G, Lotte F (2021) Guidelines to use transfer learning for motor imagery detection: an experimental study. In: 2021 10th international IEEE/EMBS conference on neural engineering (NER), pp 5–8. <https://doi.org/10.1109/NER49283.2021.9441254>. iSSN: 1948-3554
80. Yger F, Berar M, Lotte F (2016) Riemannian approaches in brain-computer interfaces: a review. IEEE Trans Neural Syst Rehabil Eng 25(10):1753–1762
81. Andreev A, Barachant A, Lotte F, Congedo M (2016) Recreational applications of OpenViBE: brain invaders and use-the-force, vol chap. 14. Wiley, London, p 241. <https://hal.archives-ouvertes.fr/hal-01366873>
82. Congedo M, Barachant A, Bhatia R (2017) Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. Brain-Computer Interfaces 4(3):155–174
83. Corsi MC, Yger F, Chevallier S, Noûs C (2021) Riemannian geometry on connectivity for clinical BCI. In: ICASSP 2021–2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 980–984. <https://doi.org/10.1109/ICASSP39728.2021.9414790>, iSSN: 2379-190X

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 10

Working with Omics Data: An Interdisciplinary Challenge at the Crossroads of Biology and Computer Science

Thibault Poinsignon, Pierre Poulain, Mélina Gallopin, and Gaëlle Lelandais

Abstract

Nowadays, generating omics data is a common activity for laboratories in biology. Experimental protocols to prepare biological samples are well described, and technical platforms to generate omics data from these samples are available in most research institutes. Furthermore, manufacturers constantly propose technical improvements, simultaneously decreasing the cost of experiments and increasing the amount of omics data obtained in a single experiment. In this context, biologists are facing the challenge of dealing with large omics datasets, also called “big data” or “data deluge.” Working with omics data raises issues usually handled by computer scientists, and thus cooperation between biologists and computer scientists has become essential to efficiently study cellular mechanisms in their entirety, as omics data promise. In this chapter, we define omics data, explain how they are produced, and, finally, present some of their applications in fundamental and medical research.

Key words Genomics, Transcriptomics, Proteomics, Metabolomics, Big data, Computer science, Bioinformatics

1 Introduction

There are different types of omics data, each revealing an aspect of cell complexity. To illustrate this complexity, we propose in Fig. 1 an analogy between the functions of a cell and that of a factory. The different omics data types are replaced there, in their specific context. Cells are the building blocks of living organisms. They can be pictured as microscopic, automated factories, made up of thousands of biological molecules (or molecular components) that work together to perform specific functions. Basically, there are four main types of molecular components: DNA, RNA, proteins, and metabolites. The whole population of one type of cellular component is named with the suffix *-ome*, i.e., *genome* (DNA), *transcriptome* (RNA), *proteome* (proteins), and *metabolome* (metabolites) (see Fig. 1). The scientific fields, which aim at studying those respective populations, are named with the suffix *-omics*,

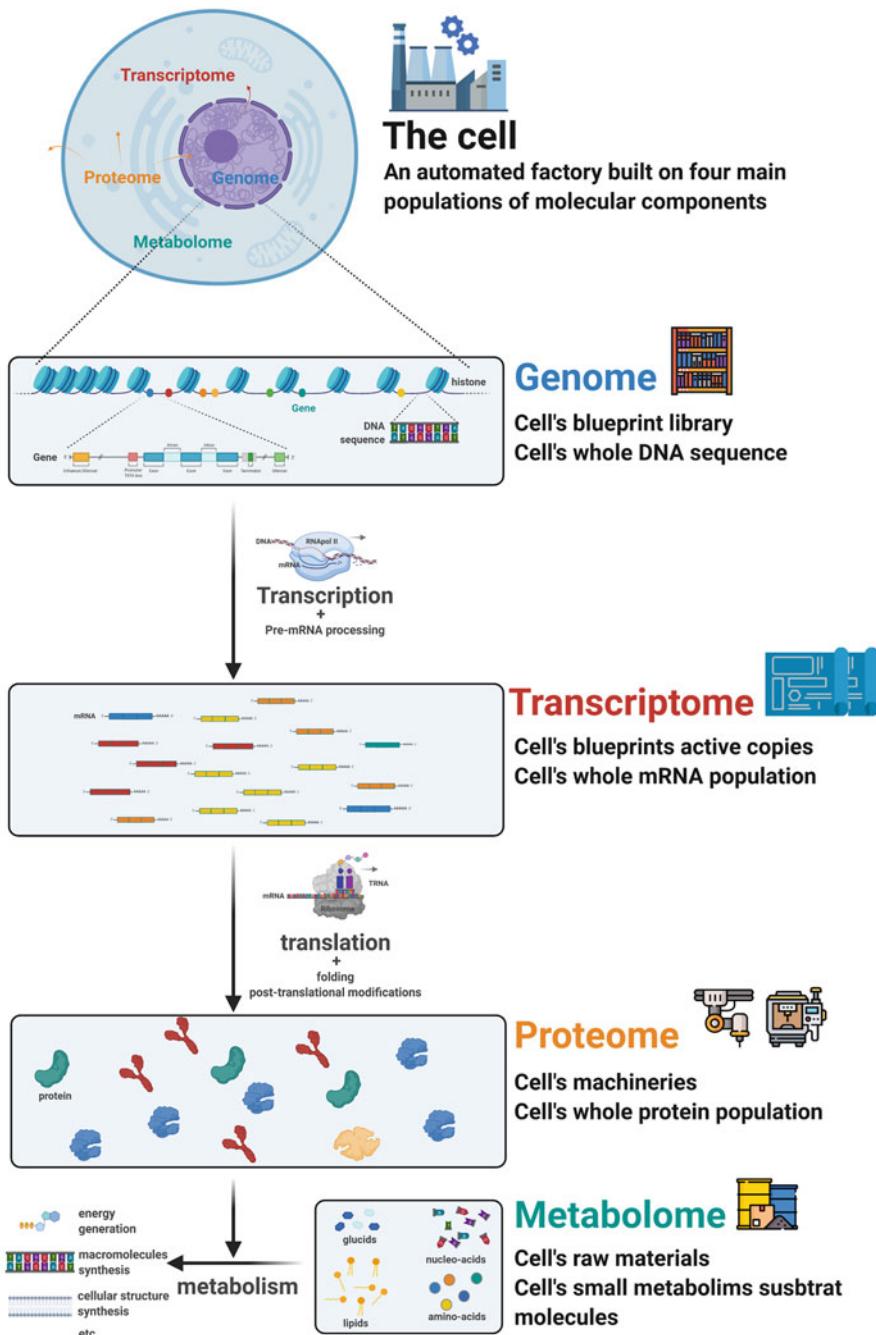


Fig. 1 The four main -omes and an analogy of their functions. The genome designates all cell's DNA molecules. The transcriptome, the proteome, and the metabolome refer, respectively, to the cell's whole set of RNA, proteins, or metabolites at a given time

i.e., *genomics*, *transcriptomics*, *proteomics*, and *metabolomics*. The common point between the different types of omics data is that they all arise from high-throughput experimental strategies that allow the simultaneous observation of all individual components that constitute either the genome, the transcriptome, the proteome, or the metabolome [1].

The genome is made of DNA molecules, which are the carrier of genetic information. It can be imagined as the blueprint library of the cell (*see* Fig. 1). From a chemical point of view, DNA molecules are polymers (or sequences) of simpler chemical units called nucleotides. There are four main types of nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G). DNA molecules are organized into chromosomes, which are compacted in the cell nucleus. The genome is directly connected to the transcriptome and the proteome (*see* next sections). The information to synthesize RNA molecules (transcriptome) and proteins (proteome) is encoded in specific regions of the DNA sequence called genes (*see* Fig. 1). Genes are made of successive nucleotides (clustered into codons), which correspond to amino acids, i.e., the molecules that constitute the proteins. The correspondence between nucleotides, codons, and amino acids is known as the genetic code. To summarize, a genomics dataset thus contains the sequences of DNA molecules present in a cell (or a population of cells) and can be seen as a copy of the cell's blueprint library (its genome) written as a long sequence of A, T, C, and G.

The transcriptome is made of RNA molecules. Multiple types exist, and they can be roughly classified into messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and non-coding RNA (ncRNA). Transcriptomics datasets mainly focus on mRNAs, which are the intermediate messengers between the genome and the proteome (*see* previous paragraph). The transcriptome is thus intimately connected to the genome and the proteome (*see* Fig. 1). Notably, the RNA polymerase is required to generate mRNA, reading the genome during transcription. In eukaryotes, mRNAs exit the nucleus to be used as templates by ribosomes (a macromolecular complex made of rRNA and proteins), to synthesize proteins by assembling amino acids (following the genetic code) during translation. Compared to the genome, the transcriptome is much more dynamic. The cell population of mRNA molecule varies according to cell requirement in proteins, and a transcriptomics dataset lists all sequences of mRNA present at a given time. They can be seen as snapshots of which parts of the genome are currently transcribed and in which proportion. Following up on the genome analogy presented in Fig. 1, mRNAs can be seen as active copies of the cell's blueprints that are more or less actively used.

The proteome is made of proteins, i.e., macromolecules made with one or several polymers of amino acids. Proteins are extraordinarily diverse in their three-dimensional (3D) conformations and associated functions. To illustrate this diversity, some proteins constitute the backbone of the cell structure, others detect or transmit external or internal chemical signals, and a large portion of them (enzymes) catalyze chemical reactions of the metabolism (the whole set of chemical reactions sustaining the cell). Proteins are also responsible for the regulation and expression (transcription and translation) of the genetic information (*see* previous paragraph). Protein functions are closely linked to their 3D spatial conformation, and all processes of the cells are based on protein activities (*see* Fig. 1). The proteome is as dynamic as the transcriptome because the set of proteins present at a given time in a cell varies accordingly to the current state and function of this cell. Proteomics datasets give a snapshot of which proteins are present at a given moment in the life of the cell. Genomics, transcriptomics, and proteomics resume the classical central dogma of biology, as first stated by Francis Crick in 1957. Even if it has been further detailed since, with, for instance, a better understanding of epigenomics, it still effectively summarizes the principal flow of information between the main molecular components of the cell: DNA is transcribed into RNA which is translated into proteins.

To end this description of omics data types, we believe it is important to mention the metabolome (*see* Fig. 1). The metabolome is made of metabolites, small molecules that are protein substrates in chemical reactions. Nucleotides and amino acids, cited before, are metabolites, as well as other molecules like lipids (forming bilayer membranes that compartmentalize the cell) or ATP (a molecule used as intracellular energy transfer). To extend, again, the analogy, metabolites can be seen as the raw materials used by the automated microscopic factory (*see* Fig. 1). Metabolomics datasets peek into the population of metabolites in a cell at a given time. Again, it is important to specify that if each cited “omics” field gives an assessment of its associated “ome” population, it is quite a “blurred” one. Everything is intertwined in a cell. Moreover, most omics studies give only an average observation on a population of cells. Multi-omics and single-cell techniques are trying to overcome these limitations.

In this chapter, we detail the different types of files used for omics data and present examples of databases where they are stored. We introduce different methods for generating omics data and finally provide some applications of omics data in fundamental research, cancer research, and pandemic response.

2 What Are Omics Data?

2.1 Results from High-Throughput Studies Written in Multiple Binary and Text Files

To describe the files used to store omics information, it is necessary to consider genomics and transcriptomics on one side and proteomics and metabolomics on the other side. Indeed, these files are generated by different experimental techniques, which are, respectively, sequencing (for genomics and transcriptomics) and mass spectrometry (for proteomics and metabolomics) (*see Fig. 2*). For each group, two types of files must be distinguished: the ones that are directly obtained after the applications of experimental protocols, i.e., the raw omics data files, and the ones that are generated by downstream informatic analyses, i.e., the processed omics data files (*see Fig. 2*). Experimental protocols and the informatic treatments applied to raw data files will be detailed in the next section.

Genomics and transcriptomics raw data files are essentially nucleotide sequence files. In that respect, the FASTA and the FASTQ text formats are commonly used. FASTA was created by Lipman and Pearson in 1985 as an input for their software [2] and became a de facto standard, without any clear statement acknowledging it [3]. This probably explains the absence of a common file extension (e.g., .fasta, .fna, .faa) even if FASTA is a unified file type. FASTA files contain one or several sequences. A sequence begins with a description line starting with the character ">". NCBI databases (*see next sections*) have unified rules to write this line.¹

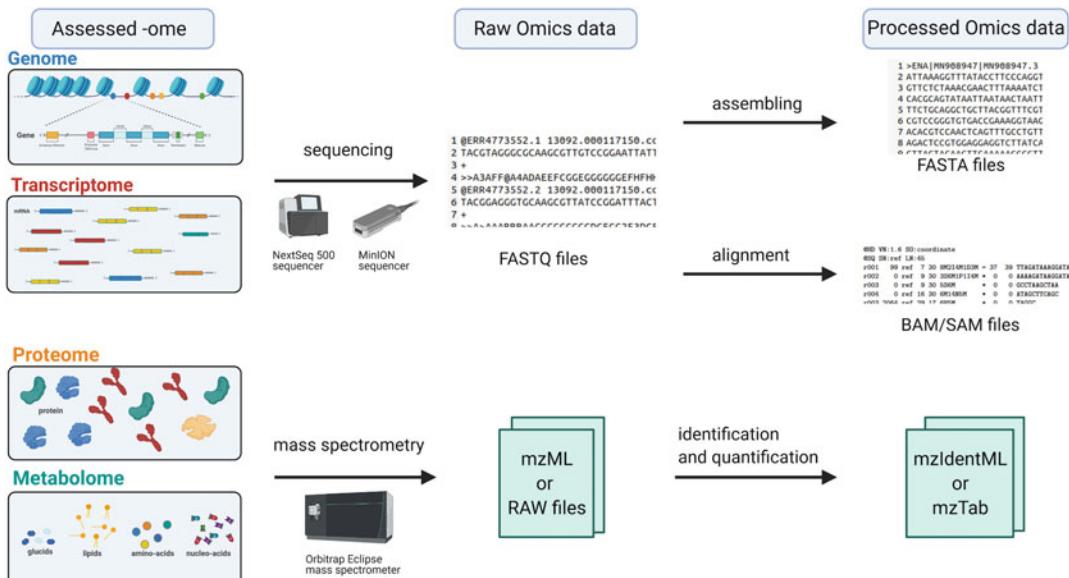


Fig. 2 Omics data are assessments of -ome populations. Raw omics data are generated through sequencing (for DNA and cDNA) or mass spectrometry (for proteins and metabolites)

¹ <https://www.ncbi.nlm.nih.gov/genbank/fastaformat/>

Subsequent lines contain the sequence itself split into multiple blocks of 60 to 80 characters (one per line). With nucleic acid sequences, the sequence lines are a series of A/T/C/G/U characters, representing the nucleic acids: adenine, thymine, cytosine, guanine, and uracil (the latter replacing thymine in RNA). FASTQ is the file format for the raw data generated by the sequencer in genomics and transcriptomics (*see* Fig. 2). The first two lines are similar as with FASTA file: identification line starts with “@” instead of “>” and the second line contains the nucleic sequence, but a quality score is associated with each position of the sequence (i.e., each letter in the sequence line). This score is called “Phred score,” and it codes the probability of error in the identification of this nucleotide [3]. It goes from 0 to 62 and is coded in ASCII symbols. This allows to code any score using a single symbol, keeping the same length as the sequence line. FASTA and FASTQ files can be opened with any text editor software. FASTQ files are mainly lists of short sequences called “reads” (between 50 and 200 nucleic acids), which need to be processed (aligned or assembled) to be further analyzed. Alignment data files are one type of processed data. Indeed, reads in FASTQ files can be aligned to a reference genome sequence to allow further analyses (*see* below for pipeline description and example of applications). The text file format used in this case is the SAM² (sequence alignment and mapping) format [4, 5]. It can be further compacted into its binary equivalent, which are BAM or CRAM formats [6].

The file formats for proteomics and metabolomics data are not as homogeneous as for genomics and transcriptomics. At least 17 types of formats exist for mass spectrometry files (*see* below) [7]. Each machine manufacturer created its own, adapted to proprietary software to read and analyze it, thus multiplying formats. In an effort to facilitate data exchange and to avoid data loss (in case of no more readable old file formats), HUPO [8] and PSI³ created the open-source mzML⁴ format (XML text file with specific tag syntax) in 2011 [9]. In the main databases that host mass spectrometry result files, most of the files are in the RAW format, developed by Thermo Fisher Scientific. These binary files contain retention time, intensity, and mass-to-charge ratios (*see* later sections). Software like Peaks, Mascot, MaxQuant, or Progenesis [10, 11] use these files to identify proteins present in the sample and to quantify them. Results from these analyses are shared through two other text file formats: mzIdentML⁵ and mzTab.⁶

² Sequence Alignment/Map Format Specification

³ HUPO Proteomics Standards Initiative

⁴ mzML 1.1.0 Specification | HUPO Proteomics Standards Initiative|mzML 1.1.0 Specification | HUPO Proteomics Standards Initiative

⁵ mzIdentML | HUPO Proteomics Standards Initiative|mzIdentML | HUPO Proteomics Standards Initiative

⁶ mzTab Specifications | HUPO Proteomics Standards Initiative|mzTab Specifications | HUPO Proteomics Standards Initiative

Note that many other file formats exist. One of the most critical for omics data analyses concerns the annotations of features on a DNA, RNA, or protein sequence. They are shared through the General Feature Format (GFF⁷) that is a text file with nine tabulated separated fields: sequence, source of the annotation, feature, start of the feature on the sequence, end of the feature, score, strand, phase, and attributes.

2.2 Results from High-Throughput Studies Shared Through Multiple Public Databases

The set of public biological databases hosting omics data is large and constantly evolving. Omics terminology started being regularly used in the 2000s. Between 1991 and 2016 (25 years), more than 1500 “molecular biology” databases were presented in publications, with a proliferation rate of more than 100 new databases each year [12]. These numbers are only the visible part of existing databases. How many have been created without being published? Around 500 of those databases are roughly co-occurrent with the apparition of the World Wide Web, the very Internet application allowing the creation of online databases. The availability of molecular biology databases decreased by only 3.8% per year from 2001 to 2016 [12]. This shows a sustained motivation from the community to create and maintain public platforms to share data. But it also highlights that this motivation comes more from a shared need for easy access to data rather than a supervised effort to coordinate approaches and unify sources. Such efforts indeed exist, for example, the ELIXIR project started in 2013 as an effort to unify all European centers and core bioinformatics resources into a single, coordinated infrastructure [13]. This notably produces the ELIXIR Core Data Resources (created in 2017), a set of selected European databases, meeting defined requirements, and the website “bio.tools,” i.e., a comprehensive registry of available software programs and bioinformatics tools. The US National Center for Biotechnology Information (NCBI⁸) databases are also main references.

Given the “raw” nature of omics dataset, they are stored in archive data repositories: raw data from scientific articles, shared on databases easily accessible for reproducibility. Except for the Sequence Read Archive (SRA), the databases cited here are mixed ones: they host raw archive data and knowledge extracted from them. For genomics dataset, NCBI database Genome [14] and EMBL-EBI (member of ELIXIR) database Ensembl [15] are references. They organize genome sequences together with annotations and include sequence comparison and visual exploration tools. Transcriptomics data can be deposited into several databases, like Gene Expression Omnibus (GEO) [16] initially dedicated to microarray datasets, which is structured into samples forming

⁷ GFF/GTF File Format

⁸ NCBI

datasets. Tools are available to query and download gene expression profiles. The Sequence Read Archive (SRA) [17] accepts raw sequencing data. PRIDE [18] is a reference database for mass spectrometry-based proteomics data. Raw files containing spectra are available with associated identification and quantification information. For metabolomics data, MetaboLights [19] is an archive data repository and a knowledge database. It lists metabolite structures, functions, and locations alongside reference raw spectra. Those databases are generalist references, and many more specialized databases exist: 89 new databases are reported in the 2021 NAR database issue, and a dozen of them are omics specific [20]. For example, AtMAD is a repository for large-scale measurements of associations between omics in *Arabidopsis thaliana*, and Aging Atlas gathers aging-related multi-omics data [21, 22]. Finally, noteworthy is the existence of general-purpose open repositories like Zenodo,⁹ which allow researchers to deposit articles, research datasets, source codes, and any other research-related digital information. Researchers thus receive credit by making their work more easily findable and reusable and hence support the application of the FAIR (findable, accessible, interoperable, reusable) data principles.¹⁰

Consistent efforts are made to cross-reference biological components (genes, proteins, metabolites) through the diversity of databases. Each database represents terabytes and petabytes of biological information (43,000 terabytes of sequence data just for SRA¹¹), and the scale of the network they form through cross-reference is hard to conceptualize. This is the “big data” in biology and even more are generated every day.

3 How to Generate Omics Data?

Genomics started in 1977 with the application of the gel-based sequencing method developed by Sanger, to sequence for the first time the whole genome of a virus: the phage phiX. Only 13 years later, in 1990, the Human Genome Project began, aiming at sequencing three billion bases of the human genome, using capillary sequencing [23]. More than 10 years and almost three billion dollars later, this titanic task was accomplished [24]. When we think of omics analyses, microarray technology remains emblematic [25]. In the 2000s, the microarray represented the keystone of a discipline then called “post-genomics” [26]. Behind this terminology, the idea was that once the genomes are entirely sequenced,

⁹ <https://zenodo.org/>

¹⁰ <https://www.go-fair.org/fair-principles/>

¹¹ NCBI Insights: The wait is over... NIH's Public Sequence Read Archive is now open access on the cloud

new studies could be performed to understand their functioning. Microarrays thus emerged as a promising tool to monitor gene expression. They allow the quantification of the abundances of transcripts, which are associated with several thousands of different genes, simultaneously. Briefly, microarrays are slides, made of glass, on which probes have been attached. These probes are small DNA molecules, which have the particularity of being specific to one (and only one) gene. The experiment then consists of extracting mRNA molecules from a population of cells and transcribing them into complementary DNA (cDNA), labeled with a fluorescent molecule. These cDNAs are then hybridized on the glass slide and end up attached to the probes which are specific to them. They create a local fluorescent signal there. The higher the amount of mRNA, the more fluorescent signal is measured at each probe location position. Microarrays have been used to successfully study many biological processes, some fundamental such as the cell cycle [27] and others directly related to health issues such as human cancer [28]. It thus paved the road to new applications for sequencing technologies (*see* below).

3.1 High-Throughput Sequencing Technologies

From 2007, new methods called next-generation sequencing (NGS) [29] helped to considerably reduce cost, technical difficulties, and duration of the process.

Illumina is the currently predominant NGS method (*see* Fig. 3). After extraction, the DNA molecules are sequenced by synthesis (SBS) on a flow cell. Thanks to sequence adaptors, each DNA molecule is amplified by bridge amplification as a cluster of copies on the flow cell. The reading of the flow cell is based on optical detection: each time a DNAPol adds a new nucleotide, a flash of light is detected. NGS advantage, compared to older Sanger techniques, is to allow massive parallel sequencing of large numbers of short sequences (between 50 and 250 nucleotides) called “reads.” The limit of this technique is the size of the fragments, but Illumina technology has very high fidelity (very low error rate).

MinION of Oxford Nanopore is another well-established NGS technology [30]. It is based on electronic detection through a nanopore (*see* Fig. 3). When there is an electric potential around a membrane (measurable as a voltage between the two sides), the passage of a macromolecule through a nanopore (a modified biological protein canal) triggers small changes in this electric potential. The changes are distinctive in function of the current nucleotide in the nanopore. So, the succession of electronic potential variation can be associated as the nucleotide sequence. This is the fundamental concept behind MinION technology, and the main advantage is the length of the sequenced molecules. Without the technical necessity of flow cells, the sequence passing through the nanopore can be very long (order of magnitude of a thousand instead of a hundred base pairs) [31]. But given that the physical

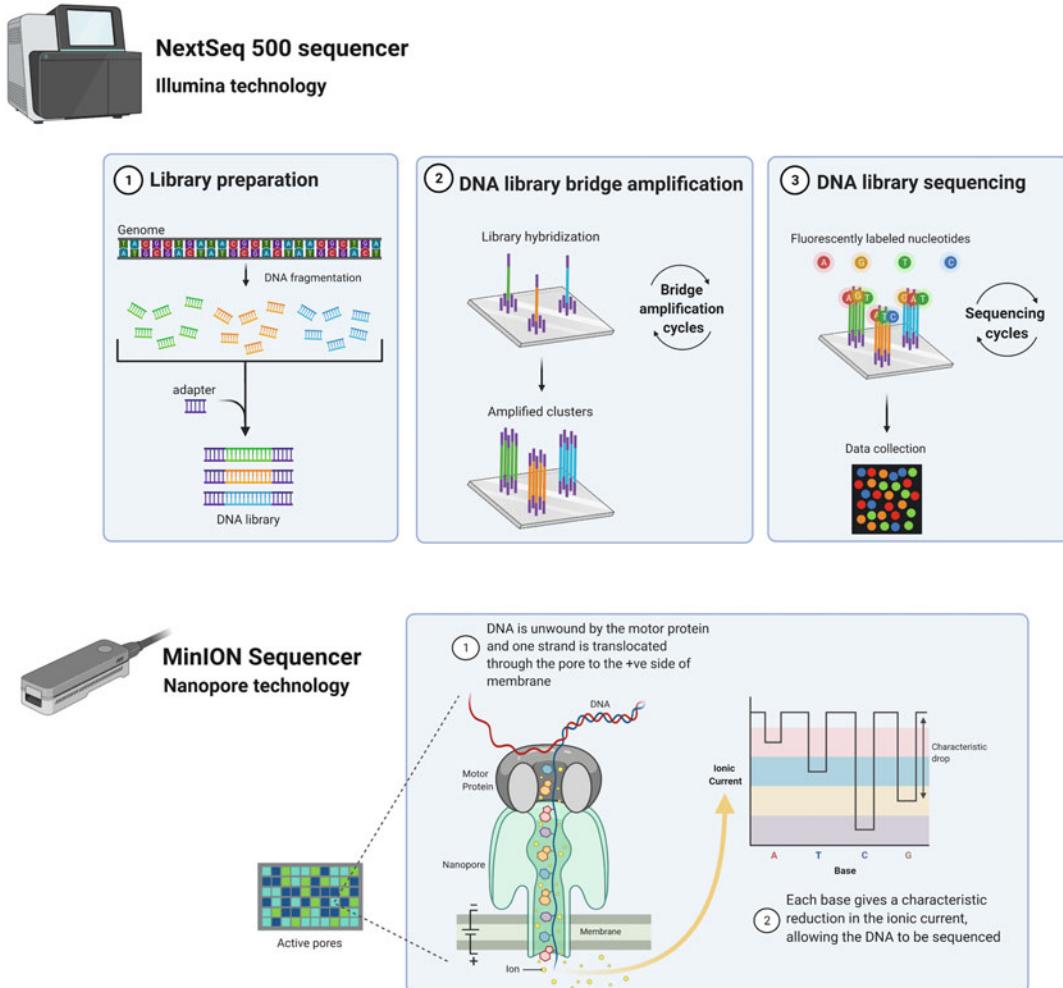


Fig. 3 Illumina and MinION sequencing technologies. Illumina is a sequencing by synthesis technology that allows massive parallel sequencing of small DNA molecules. MinION is a nanopore-based technology that allows the sequencing of longer DNA molecules

signal detected is small variations of an electric potential, the sequencing is less reliable (higher error rate). Depending on the fidelity of the sequencing or the size of the sequence needed, SBS and nanopore-based techniques are complementary.

The sequencing machine output is a group of FASTQ files (*see* previous section). For genomics data, fragments must be assembled to obtain a single sequence of the genome. For transcriptomics data, fragments can be aligned on a reference genome to observe which genes are transcribed at a given time (transcriptome *de novo* assembly is also possible but still very challenging). Therefore, to extract information from the FASTQ files produced by the sequencer, two main processing steps are needed. The numerous small sequences (reads) stored in the file must be aligned to a

reference genome (mapping), and then the count of reads aligned to a gene sequence gives an estimation of its level of transcription (quantification). Dozens of bioinformatics tools have been developed over the years for mapping (STAR [31], TopHat [32], HISAT2, Salmon [33]) and quantification (featureCounts [34], Cufflinks [35]). Benchmarking studies highlight similar performance for most of them [36–38]. Interestingly, TopHat2 exhibits an alignment recall on simulated malaria data that varies from under 3% using defaults to over 70% using optimized parameters [39]. This underlines the impact of parameter optimization on result quality. Quantification tools generate a text file summarizing the level of transcription of each gene in each condition into a matrix of counts.

3.2 Mass Spectrometry Technologies

Since the first use of a mass spectrometer for protein sequencing in 1966 by Biemann,¹² the improvement of mass spectrometer is closely linked to proteomics and metabolomics development [40]. Metabolites and proteins cannot be read as templates like DNA or RNA, and so they neither can be amplified nor sequenced by synthesis. To access their sequence, the main tool is the mass spectrometer. In the classical bottom-up approach, proteins are digested into small peptides, which pass through a chromatography column. They are then sequentially sprayed as ions into the spectrometer. Migration through the spectrometer allows separation of the peptides according to their mass-to-charge ratio. For each fraction exiting the column, an abundance is calculated. In a data-dependent acquisition (DDA), a few peptides with an intensity superior to a given threshold are isolated one at the time. They are fragmented, and additional spectra (mass-to-charge ratio and intensity) are generated for each fragmented ion. In a data-independent acquisition (DIA), a spectrum is generated for all fractions coming out of the chromatography column. Obtained spectra are a combination of spectra corresponding to each peptide present in each original fraction. Comparison with a peptide spectrum library generated in silico is therefore required to allow the deconvolution of those complex spectra. All this information (abundances in fractions, mass-to-charge ratios, intensities) is stored into .raw files, which can only be read by dedicated software (*see Subheading 2.1*).

3.3 Single-Cell Strategies

Most omics experiments are bulked, and they are an average measure done on a population of cells, which is more or less homogeneous. Single-cell omics allow a more precise measurement, highlighting the plasticity of the cell system. Single-cell techniques started with manual separation of a single cell under a microscope in 2009 [41] and quickly evolved toward techniques allowing the

¹² HUPO—Proteomics Timeline

parallel sequencing of thousands of cells [42]. Plate-based techniques use flow cytometry to separate isolated cells into the different wells of a plate, allowing processing of hundreds of cells. The introduction of nanometric droplets to separate isolated cells allowed the parallel processing of thousands of cells thanks to individual barcoding [43, 44]. Cells isolated from tissues are mixed with microparticles in a buffer that forms droplets in oil. Most droplets are empty, but some contain both a microparticle and a cell. After cell lysis, oligonucleotide primers on the microparticles allow the capture of the cell mRNA (by oligo-dT and polyA tail complementarity). Primers on the same microparticle are barcoded, thus creating a cell tag on each sequence. Amplification and sequencing can be bulked without losing the cell of origin for each transcript. Several bioinformatics tools are specialized for single-cell transcriptomics data [45]. For example, Cell Ranger and Loupe Browser are, respectively, four pipelines (mapping, quantification, and downstream analysis) and a visualization tool developed by 10 \times Genomics [44]. Single-cell transcriptomics data are challenging for bioinformatics analysis because of their high level of technical noise and the multifactorial variability between cells [45]. Transcriptomics is the more advanced single-cell omics, but single-cell genomics is also used in SNP and copy number variation screening (*see* Subheading 4.2).

Proteomics and metabolomics data are still challenging to obtain at a single cell level: one cell yields only 250–300 pg [46] of proteins when MS in-depth measurement still necessitates population scale yield. But thanks to innovations in sample preparation and experimental design, single-cell proteomics assessments scaled up from a few hundred to more than a thousand identified proteins in just 4 years [47].

4 Which Applications for Omics Data?

4.1 In Fundamental Research

Describing biological systems implies to identify, quantify, and functionally connect their individual molecular components. Given the diversity of cellular components and their multiple interlocking functions, the large scale of omics data empowers the characterization of biological systems. As stated before, each type of “omics” is an assessment of a specific subpopulation of molecular components. Mining omics data thus allows bulk identification of the nature (sequence and structure), location, function, and abundance of molecular components in those subpopulations.

Genomics data are making the genome sequences of thousands of species accessible. The first direct application of these resources is the annotation of genomic features onto those genomic sequences: protein-coding genes, tRNA and rRNA genes, pseudogenes, transposons, single-nucleotide polymorphisms, repeated regions, telomeres, centromeres... Genomic features are numerous, and DNA sequences alone can be enough to recognize

patterns specific to some of them. For example, specific tools exist to detect protein-coding genes, like Augustus¹³ [48]. The annotation can be based only on sequence patterns or also on comparison with another sequence. Comparative genomics, i.e., the comparison of genome sequences, allows the transfer of knowledge for homolog genes (evolutionarily related genes) between species. Bioinformatics tools exist to infer evolutionary relationships between genes based on their sequence similarity [49]. Understanding the evolution of the genome helps to understand the dynamics behind phenotypic convergence, population evolutions, speciation events, and natural selection processes. For example, the study of 17 marine mammals' genomes offered insight into the macroevolutionary transition of marine mammal lineages from land to water [50].

Transcriptomics data give insight on the levels of gene transcription. The resulting count matrix (*see* previous section) is mainly used to carry out differential expression analysis (DEA) of genes between conditions. Conditions differ by the variation of a single factor: a mutation, a different medium, or a stimulus. Basic DEA is a multi-step workflow [51] that allows the detection of statistically significant variations in expression across conditions. The final goal is to deduce insight on the gene's functions from the observed variations. Transcriptomics data are also used to increase the quality of genome annotation. The presence of hypothetical genes can be verified by their transcription, the exact structure of known genes can be refined (size of UTRs and exons; *see* Fig. 1), and previously undetected genes can be observed [52].

Proteomics data allows the identification and quantification of proteome. Proteome does not totally correlate with transcriptome. RNA can be spliced (assembly of the mRNA from exons, not always the same and in the same order), and proteins undergo several post-translational modifications (minor changes in the chemical structure of the protein) and re-localization [53]. Cellular pathways and phenotypes thus cannot be fully understood only through transcriptomics assessments. Proteomics completes the information given by genomics and transcriptomics. It describes the third -ome of the central dogma of biology (*see* Fig. 1).

Multi-omics analysis, taking advantage of several omics insights in the same experimental approach, comes with several challenges. Generating several types of omics data comes with a significant investment in time, skilled manpower, and money [1]. Even if generated in the same experimental approach, omics data are heterogeneous by nature, thus complexifying their integration. If challenging, multi-omics datasets are also a step toward the systemic description of biological systems [54].

¹³ Augustus/ABOUT.md at master

4.2 In Medical Research

An early application of genomics in medical research is the genome-wide association studies (GWAS). By comparing genome sequences from a large population of individuals (both healthy and sick), GWAS highlight SNPs (single-nucleotide polymorphisms) that are significantly more frequent in individuals with the disease. Correlation does not mean causality, but GWAS can give a first clue of the metabolic pathways or cellular components involved in the disease [55]. This strategy has proven to be efficient in the case of “common complex diseases.” Unlike Mendelian diseases (which are rarer), the heritability (genetic origin) of these diseases depends on hundreds of SNPs with small effect sizes, which GWAS studies help identify [56]. Alzheimer’s disease and cancers are examples of “common complex diseases” whose genetic underpinnings have been explored through GWAS [55, 57].

Most cancers emerge from the successive alteration of cell functioning (by accumulation of mutations), leading to abnormal growth causing tumors and metastasis. Multi-omics studies can highlight the underlying molecular mechanisms of cancer development, better explain resistance to treatment, and help classify cancer types. Screening cohorts of patients helps assess alleles associated with the development of certain types of cancer. The different subtypes for breast cancer are a well-documented example [58].

Single-cell genomics is the only way of characterizing rare cellular types such as cancer stem cells [59]. Single-cell omics data are also used to follow the rapid evolution of cancer cell population inside tumors. Understanding and describing cancer cell population dynamics is crucial: the characteristic accelerated rate of mutation can be the cause of treatment resistance. Omics data specific to cancer cell lines are shared on specific databases driven and maintained by global consortium such as the Cancer Genome Atlas Program¹⁴ (over 2.5 petabytes of genomics, epigenomics, transcriptomics, and proteomics data) or the International Cancer Genome Consortium [60].

Omics data proved to be a priceless resource in pandemic response. The virus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causing the COVID-19 disease quickly spread around the world, causing more than six million deaths (as of March 2022) and a global health crisis. Its RNA sequence was obtained in January 2020 and allowed the development of detection kits and later RNA-based vaccines. Since the beginning of the pandemic, the genomic evolution of the virus is followed almost in real time, as new variants (with mutations affecting mostly the spike protein of the virus envelope) are sequenced. Variant profiling allows the World Health Organization to closely monitor variants of concern. The precise characterization of the virus structure

¹⁴ <https://www.cancer.gov/tcga>

opens the research of therapeutic targets. Multi-omics studies helped specify the COVID-19 biomarkers, pathophysiology, and risk factors [61].

Getting omics data in brain tissue studies is promising but challenging because of brain specificity. Indeed, except in a few specific diseases where *in vivo* resections are performed (brain tumors, surgically treated epilepsy, etc.), human brain samples are collected postmortem, when the less stable molecule populations are already significantly altered. For example, studies of the brain transcriptome are deeply impacted. On the other hand, some omics studies target peripheral fluids (e.g., plasma, cerebrospinal fluid, etc.) with the aim to find biomarkers, but the relationships between observations in peripheral fluids and pathophysiological mechanisms in the brain are far from clear. Moreover, the brain is organized as a network of intricate substructures, constituted of several cell types (glial cells and different neuron types) with distinct function and thus different omics landscape [62]. Nonetheless, multi-omics exploratory studies are describing complex diseases in a systematic paradigm, highlighting diversity of cellular dysregulations linked to complex pathologies like Alzheimer's disease [57].

5 Conclusion

Genomics, transcriptomics, proteomics, and metabolomics are arguably the most developed and used omics, but they are not the only ones. Other omics describe other sides of the functioning of the cell, which require intricate relationships between omics levels. For example, epigenomics describes the transitory chemical modifications of DNA, and lipidomics looks at the lipidic subpopulation of metabolites (*see* Fig. 1). Omics diversity mirrors the complexity of cell systems. With the constant improvement of measurement techniques, possibilities to assess ever larger subsystems of the cells are increasing. Omics dataset generation is paired with the development of software, essential tools to generate, read, and analyze them. By design, computer science is therefore omnipresent in modern “big data” biology. The need for more gold standard analysis pipelines and file formats grows with the scale and complexity of produced datasets.

Acknowledgments

This work was funded by the Agence Nationale pour la Recherche (MINOMICS, grant number ANR-19-CE45-0017).

The authors are grateful to Sarah Cohen-Boulakia for reviewing this chapter.

Figures were made on [BioRender](#), using icons from [Flaticon](#).

References

1. Hasin Y, Seldin M, Lusis A (2017) Multi-omics approaches to disease. *Genome Biol* 18:83
2. Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. *Science* 227:1435–1441
3. Cock PJA, Fields CJ, Goto N et al (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* 38:1767–1771
4. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079
5. Danecek P, Bonfield JK, Liddle J et al (2021) Twelve years of SAMtools and BCFtools. *GigaScience* 10:giab008
6. Hsi-Yang Fritz M, Leinonen R, Cochrane G et al (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res* 21:734–740
7. Deutsch EW (2012) File formats commonly used in mass spectrometry proteomics. *Mol Cell Proteomics* 11:1612–1621
8. Deutsch EW, Lane L, Overall CM et al (2019) Human proteome project mass spectrometry data interpretation guidelines 3.0. *J Proteome Res* 18:4108–4116
9. Martens L, Chambers M, Sturm M et al (2011) mzML—a community standard for mass spectrometry data. *Mol Cell Proteomics* 10(R110):000133
10. Ma B, Zhang K, Hendrie C et al (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 17:2337–2342
11. Välikangas T, Suomi T, Elo LL (2017) A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Brief Bioinform*
12. Imker HJ (2018) 25 years of molecular biology databases: a study of proliferation, impact, and maintenance. *Front Res Metr Anal* 3:18
13. Harrow J, Drysdale R, Smith A et al (2021) ELIXIR: providing a sustainable infrastructure for life science data at European scale. *Bioinformatics* 37:2506–2511
14. Benson DA, Karsch-Mizrachi I, Lipman DJ et al (2010) GenBank. *Nucleic Acids Res* 38: D46–D51
15. Howe KL, Achuthan P, Allen J et al (2021) Ensembl 2021. *Nucleic Acids Res* 49:D884–D891
16. Barrett T, Wilhite SE, Ledoux P et al (2012) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 41: D991–D995
17. Leinonen R, Sugawara H, Shumway M et al (2011) The sequence read archive. *Nucleic Acids Res* 39:D19–D21
18. Perez-Riverol Y, Csordas A, Bai J et al (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 47: D442–D450
19. Haug K, Cochrane K, Nainala VC et al (2019) MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res* gkz1019
20. Rigden DJ, Fernández XM (2021) The 2021 *Nucleic Acids Research* database issue and the online molecular biology database collection. *Nucleic Acids Res* 49:D1–D9
21. Lan Y, Sun R, Ouyang J et al (2021) AtMAD: *Arabidopsis thaliana* multi-omics association database. *Nucleic Acids Res* 49:D1445–D1451
22. Aging Atlas Consortium, Liu G-H, Bao Y et al (2021) Aging Atlas: a multi-omics database for aging biology. *Nucleic Acids Res* 49:D825–D830
23. Karger BL, Guttmann A (2009) DNA sequencing by CE. *Electrophoresis* 30:S196–S202
24. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431: 931–945
25. Schena M, Shalon D, Davis RW et al (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–470
26. Gershon D (1997) Bioinformatics in a post-genomics age. *Nature* 389:417–418
27. Spellman PT, Sherlock G, Zhang MQ et al (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 9:25
28. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 14(4):457–460
29. Metzker ML (2010) Sequencing technologies — the next generation. *Nat Rev Genet* 11:31–46

30. Deamer D, Akeson M, Branton D (2016) Three decades of nanopore sequencing. *Nat Biotechnol* 34:518–524
31. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21
32. Kim D, Pertea G, Trapnell C et al (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14:R36
33. Patro R, Duggal G, Love MI et al (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 14: 417–419
34. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923–930
35. Trapnell C, Roberts A, Goff L et al (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7:562–578
36. Teng M, Love MI, Davis CA et al (2016) A benchmark for RNA-seq quantification pipelines. *Genome Biol* 17:74
37. The RGASP Consortium, Engström PG, Steijger T et al (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* 10:1185–1191
38. Schaarschmidt S, Fischer A, Zuther E et al (2020) Evaluation of seven different RNA-Seq alignment tools based on experimental data from the model plant *Arabidopsis thaliana*. *Int J Mol Sci* 21:1720
39. Baruzzo G, Hayer KE, Kim EJ et al (2017) Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat Methods* 14: 135–139
40. Biemann K, Tsunakawa S, Sonnenbichler J et al (1966) Structure of an odd nucleoside from serine-specific transfer ribonucleic acid. *Angew Chem Int Ed Engl* 5:590–591
41. Tang F, Barbacioru C, Wang Y et al (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6:377–382
42. Svensson V, Vento-Tormo R, Teichmann SA (2018) Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* 13: 599–604
43. Macosko EZ, Basu A, Satija R et al (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161:1202–1214
44. Zheng GXY, Terry JM, Belgrader P et al (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 8:14049
45. Stein CM, Weiskirchen R, Damm F et al (2021) Single-cell omics: overview, analysis, and application in biomedical science. *J Cell Biochem* 122:1571–1578
46. Jehan Z (2019) Single-cell omics: an overview. In: *Single-cell omics*. Elsevier, pp 3–19
47. Kelly RT (2020) Single-cell proteomics: progress and prospects. *Mol Cell Proteomics* 19: 1739–1748
48. Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 33:W465–W467
49. Quest for Orthologs consortium, Altenhoff AM, Boeckmann B et al (2016) Standardized benchmarking in the quest for orthologs. *Nat Methods* 13:425–430
50. Yuan Y, Zhang Y, Zhang P et al (2021) Comparative genomics provides insights into the aquatic adaptations of mammals. *Proc Natl Acad Sci* 118:e2106080118
51. Van den Berge K, Hembach KM, Soneson C, Tiberti S, Clement L, Love M, Patro R, Robinson MD (2019) RNA sequencing data: Hitchhiker's guide to expression analysis. *Annu Rev Biomed Data Sci* 2:139–173
52. Chen G, Shi T, Shi L (2017) Characterizing and annotating the genome using RNA-seq data. *Sci China Life Sci* 60:116–125
53. Arivaradarajan P, Misra G (eds) (2018) Omics approaches, technologies and applications: integrative approaches for understanding OMICS data. Springer Singapore, Singapore
54. Veenstra TD (2021) Omics in systems biology: current progress and future outlook. *Proteomics* 21:2000235
55. Tam V, Patel N, Turcotte M et al (2019) Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 20:467–484
56. Uitterlinden A (2016) An introduction to genome-wide association studies: GWAS for dummies. *Semin Reprod Med* 34:196–204
57. Hampel H, Nisticò R, Seyfried NT et al (2021) Omics sciences for systems biology in Alzheimer's disease: state-of-the-art of the evidence. *Ageing Res Rev* 69:101346
58. Kohler BA, Sherman RL, Howlader N et al (2015) Annual report to the nation on the status of cancer, 1975–2011, featuring

- incidence of breast cancer subtypes by race/ethnicity, poverty, and state. *JNCI J Natl Cancer Inst* 107
59. Liu J, Adhav R, Xu X (2017) Current progresses of single cell DNA sequencing in breast cancer research. *Int J Biol Sci* 13:949–960
60. Zhang J, Bajari R, Andric D et al (2019) The international cancer genome consortium data portal. *Nat Biotechnol* 37:367–369
61. Overmyer KA, Shishkova E, Miller IJ et al (2021) Large-scale multi-omic analysis of COVID-19 severity. *Cell Syst* 12:23–40.e7
62. Naumova OY, Lee M, Rychkov SY et al (2013) Gene expression in the human brain: the current state of the study of specificity and spatio-temporal dynamics. *Child Dev* 84:76–88

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 11

Electronic Health Records as Source of Research Data

Wenjuan Wang, Davide Ferrari, Gabriel Haddon-Hill, and Vasa Curcin

Abstract

Electronic health records (EHRs) are the collection of all digitalized information regarding individual's health. EHRs are not only the base for storing clinical information for archival purposes, but they are also the bedrock on which clinical research and data science thrive. In this chapter, we describe the main aspects of good quality EHR systems, and some of the standard practices in their implementation, to then conclude with details and reflections on their governance and private management.

Key words Electronic health records, Data science, Machine learning, Data quality, Coding schemes, SNOMED-CT, ICD, UMLS, Data governance, GDPR

1 Introduction

Vast quantities of data are routinely recorded as part of the care process. While its primary aim is managing individual's patient care, there are significant opportunities to use these data to address research questions of interest. In the United Kingdom, there has been almost 25 years of research using routine primary care data, anonymized at source, through the General Practice Research Database (now CPRD, Clinical Practice Research Datalink [1]), and other data sources, also pooling data from multiple practices and tied to specific electronic health record (EHR) systems (QResearch [2], ResearchOne [3]). As better described in Subheading 4, we define anonymized data as one for which all elements that can link back to its owner are irrecoverably deleted; alternately there are pseudo-anonymization options that allow the reidentification of the owner through a procedure mediated by those responsible for that data security and privacy protection. Health Data Research UK has created a nationwide registry of EHR-derived datasets available for research [4]. A similar development has taken place in the Netherlands, where, in the early 1990s, the Netherlands Institute for Health Services Research (NIVEL) developed its Netherlands Information Network of General Practice [5], now named NIVEL

Primary Care Database (NIVEL-PCD) [6, 7]. Belgium also has its Intego Network [7, 8]. France has the *Système National des Données de Santé* [9, 10] and the data warehouse of *Assistance Publique-Hôpitaux de Paris (AP-HP)* [11]. Sweden has numerous and extensive nationwide registries [12]. These databases provide valuable information about the use of health services and developments in population health. In the United States, there has not been a tradition of using routine anonymized data, largely because the Health Insurance Portability and Accountability Act (HIPAA) regulations place restrictions on the linkage of health data from different sources without consent [13–15] and because small office practices have not been widely computerized. Instead, the focus has been mainly on secondary care (hospital) data, facilitated by the National Institutes of Health's (NIH) Clinical Translational Science Awards (CTSA) [16]. Use or reuse of administrative data for research purposes is becoming more restricted in Europe as well, partly as a consequence of the European General Data Protection Regulation (GDPR) that was established in 2016 [17, 18]. In addition, data owners increasingly want control over the use of their data, making it more difficult to construct large centralized databases.

2 Data Quality in EHR

An electronic health record (EHR) is a digital version of a patient's medical history which may include all of the key administrative clinical data relevant to that person's care, including demographics, vital signs, diagnoses, treatment plans, medications, past medical history, allergies, immunizations, radiology reports, and laboratory and test results. EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users. EHRs have been adopted with the aim of improving quality of patient care quality, in particular by ensuring that all pertinent medical information is being shared as needed for different care providers. Meantime, the rapidly growing number of EHRs has led to increasing interest and opportunities for various research purposes. To ensure the patients receive care as they need and to draw valid and reliable research findings, quality data are needed.

Data quality is defined as “the totality of features and characteristics of a data set that bear on its ability to satisfy the needs that result from the intended use of the data” [19]. Currently, there is no definitive agreement on the components of data quality in available research. Feder described in a study [20] frequently reported components of data quality including data accuracy (data must be correct and free of errors), completeness (data must be sufficient in breadth, depth, and scope for its desired use), consistency (data must be presented in a consistent format),

credibility (data must be regarded as true and credible), and timeliness (data should be recorded as quickly as possible and used within a reasonable time period) [20–24]. Other aspects of data quality might include accessibility which means that data must be available for use or easily retrievable, appropriate amount of data which means the quantity of data must be appropriate, ease of understanding which means data must be clear, interpretability which means data must be in appropriate language and units, etc.

Many concerns were raised on digital data quality within EHRs including incompleteness, duplication, inconsistent organization, fragmentation, and inadequate use of coded data within EHR workflows [25]. As the old programming maxim states: garbage in, garbage out. Poor data quality can impact the care patients receive which may in turn lead to long-term damage or even death. It will also impact public health decision-making whenever it is based on statistics drawn from inaccurate data. In the following section, we will investigate in more detail the challenges regarding data accuracy and data completeness.

2.1 Data Accuracy

Data accuracy can be conceptualized as how accurate or truthful the data captured through the EHR system is. In other words, it is the degree to which the value in the EHR is a true representation of the real-world value [20, 23, 24] (e.g., whether a medication list accurately reflects the number, dose, and specific drugs a patient is currently taking [21]). A pilot study evaluated information accuracy in a primary care setting in Australia and confirmed that errors and inaccuracies exist in EHR [26]. This pilot study showed that high levels of accuracy were found in the area of demographic information and moderately high levels of accuracy were reported for allergies and medications. A considerable percentage of non-recorded information was also present. The sources of data inaccuracy could be mistakes made by clinicians (e.g., clinicians improperly use the “cut and paste” function in electronic systems [27]), error, loss or destruction of data during a data transfer [27]. Ways to improve data accuracy at collection include avoiding EHR pitfalls (e.g., fine-tuning preference lists, being careful when copying data, modifying templates as needed, documenting what was done, etc.) and being proactive (e.g., conducting regular internal audits, training staff, maintaining a compliance folder, etc.).

Data accuracy can be assessed via different approaches [20]. One can compare a given variable within the dataset to other variables which is referred to as internal validity, e.g., using medication to confirm the status of the disease. Internal validation can also be done by looking for unrealistic values (a blood pressure that is too high or low [28]) which could be checked by identifying outliers. One can also use different data sources or datasets to cross-check the data accuracy which is referred to as external validity, e.g., a patient was registered in a stroke registry but recorded as not

having a stroke in the current dataset. Generally, it is hard to link multiple datasets due to data privacy policy. Simple statistical measures can help the researcher determine whether variable values follow logical restrictions and patterns in the data such as central tendency (e.g., mean, median, mode) and dispersion (range, standard deviation) for continuous variables and frequencies and proportions for categorical variables and goodness-of-fit tests (e.g., Pearson chi-square) [20]. Researchers found that validation helps check the quality of the data and identify types of errors that are present in the data [28].

2.2 Data Completeness

Data completeness is referred to as the degree and nature of the absence of certain data fields for certain variables or participants. Generally, these absent values are called missing data. Missing data is very common in all kinds of studies, which can limit the outcomes to be studied, the number of explanatory factors considered, and even the size of the population included [28] and thus reduce the statistical power of a study and produce biased estimates, leading to invalid conclusions [29]. Data may be missing due to a variety of reasons. Some data might not be collected due to the design of the study. For example, in some questionnaires, certain questions are only for females to answer which leads to a blank for males for that question. Some data may be missing simply because of the breakdown of certain machines at a certain time. Data can also be missing because the participant did not want to answer. Some data might be missing due to mistakes during data collection or data entry. Thus, knowing how and why the data are missing is important for subsequent handling and for analyzing the mechanism underlying missing data.

Depending on the underlying reason, missing data can be categorized into three types [30] (Fig. 1): missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR is defined as data to be missing not related to any other variables or the variable itself. Examples of MCAR are failures in recording observations due to random failures with experimental instruments. The reasons for its absence are normally external and not related to the observations themselves. For MCAR, it is typically safe to remove observations with missing values. The results will not be biased but the test might not be powerful as the number of cases is reduced. This assumption is unrealistic and hardly happens in practice. For missing data that are MAR, missingness is not random and can be related to the observed data but not to the value of this given variable [31]. For example, a male participant may be less likely to complete a survey about depression severity than a female participant [32]. The data is missing because of gender rather than because of the depression severity itself. In this case, the results will be biased if we remove patients with missing values as most completed observations are

MCAR		
Definition: Data to be missing is not related to any other variables or the variable itself.	Example: Failures in recording observations due to random failures with experimental instruments	Acceptable approaches: - Complete case analysis - Single imputation - Multiple imputation
MAR		
Definition: Missingness is not random and can be related to the observed data but not to the value of this given variable.	Example: A male participant may be less likely to complete a survey about depression severity than a female participant.	Acceptable approaches: - Complete analysis - Pairwise deletion - Regression or likelihood based - Multiple imputation
MNAR		
Definition: Missingness is related to the value that the missing data would have had.	Example: The participant refuses to report their depression severity because they are seriously depressed.	Acceptable approaches: - Joint modelling of the outcome as well as the relation between outcome and probability of response

Fig. 1 Summary on missing mechanisms with definitions, examples, and acceptable approaches for handling the missing values

females. Thus, other observed variables of the participants should be accounted for properly when imputing missing data that are MAR. But MAR is an assumption that is impossible to verify statistically [33] and substantial explorations and analysis are needed. MNAR refers to situations where missingness is related to the value that the missing data would have had. For example, the participant refuses to report their depression severity because they are seriously depressed. In this case, missingness is due to the value itself and no other data can predict this value. Missing data that are MNAR are more problematic as one may lack data from key subgroups which, in turn, may lead to samples that are not representative of the population of interest. The only way to obtain an unbiased estimate of the parameters in such a case is to model the missing data and then be incorporated into a more complex one for estimating the missing values [29].

Handling missing data is critical and should be done according to the assumption on the missingness mechanism, as the results might be biased if handled differently. Techniques for handling missing data include the following [29]:

1. Complete case analysis (also known as listwise deletion) to simply omit those cases with the missing data. This approach is suitable for MCAR assumption or when the level of missingness is low in a large dataset.

2. Pairwise deletion allows researchers to use cases with missing values but the variable with missing values will not be included in the analysis. This method is known to be less biased for MCAR or MAR data [29]. The analysis will be deficient if there is a high level of missingness in the data [29].
3. Single imputation means that missing values are replaced by a value defined by a certain rule. Here is a list of possible imputation rules. (1) A simple imputation rule is to substitute the missing value with the mean, median, or mode. (2) A more sophisticated approach uses regression (the missing values are predicted from the other variables using regression). (3) Last observation carried forward or next observation carried backward is for longitudinal data (i.e., repeated measures). If a certain measure is missing, the previous observation or the next observation can be used to impute the current missing values. (4) Maximum likelihood method assumes that the observed data are a sample drawn from a multivariate normal distribution and the missing data are imputed with the maximum likelihood method [34]. (5) K-nearest neighbors method can be used to impute the missing values with the average from the k-nearest neighbors. Single imputation often results in an underestimation of the variability since the unobserved value is analyzed as the known, observed values [35] and some single imputation methods depend on specific rules (e.g., last observation carried forward) rather than missing mechanism assumption which are often unrealistic [36]. Single imputation is often a potentially biased method and should be used with great caution [35–38].
4. Multiple imputation consists in replacing missing values with a set of plausible values which contain the natural variability and uncertainty of the correct values [29]. The multiple imputed values are predicted using the existing data from other variables [39], and then multiple imputed datasets are generated using the set of values. Compared to single imputations, creating multiple imputations accounts for the statistical uncertainty in the imputations. A typical method for multiple imputation is the use of chained equations (MICE) [40]. Multiple imputation operates under the assumption that the missing data are MAR since we use other variables to predict the missing values. Implementing MICE when data are not MAR could result in biased estimates [40]. Multiple imputation has been shown to be a valid method for handling missing data and is considered a good approach for datasets with a large amount of missing data. This method is available for most types of data [31, 37, 38]. Studies comparing software packages for multiple imputations are available [41].

The acceptable handling methods for different missing data mechanisms [35] are summarized in Fig. 1. For MCAR, the methods for handling missing data which give unbiased effects and standard errors are complete case analysis, regression or likelihood-based single imputation methods, and multiple imputation. For MAR assumption, pairwise deletion, regression or likelihood-based single imputation methods, and multiple imputation provide unbiased effects. Under the MNAR assumption, the above methods are no longer suitable. In this case, the appropriate analysis requires the joint modeling of the outcome along with the missing data mechanism [35]. This could be done by asking related questions, e.g., (1) what's the probability of having missing data given the outcome and (2) what's the probability of an outcome in those with missing data? Selection [33] and pattern-mixture models [42] are two example approaches for modeling the above two questions, respectively.

The recommended strategies to overcome barriers caused by missing data would be to first understand the data and the missing mechanism. If the data are simply unavailable, alternative datasets and similar information might be available [28]. Then the imputation method could be selected based on the understanding of the missing values. Since the correctness of the assumptions cannot be definitively validated, it is recommended to perform a sensitivity analysis to evaluate the robustness of the results to the deviations from the assumptions [28].

2.3 Other Challenges and General Practices Recommendations

There are other challenges in EHR data. For example, some data may be recorded without specifying units of measurement which makes these data hard to interpret [28]. In this case, an understanding of the data collection process and background knowledge can be helpful in interpreting the data. There might be inconsistencies in data collection and coding across institutions and over time [28]. Some inconsistencies can be easily identified from the data, e.g., a measure was started to be recorded only after a certain time. On the other hand, some inconsistencies may be hard to identify and require an understanding of how data are collected geographically and over time. Last but not least, unstructured text data residing in the EHR causes poor accessibility and other data quality issues such as a lack of objectivity, consistency, or completeness [28]. Data extraction techniques such as natural language processing (NLP) are being used to identify information directly from text notes.

Quality data is the basis for a valid research outcome and whether the quality is enough depends on the purpose of the study. Currently, there are no certain criteria for deciding whether the quality of the data is sufficient, but careful analysis of the data quality should help the researchers decide if the data at hand is useful for the study [28]. Three general practices were

recommended by Feder [20]. The first recommendation is to get familiar with the EHR platform and EHR-based secondary data source. Knowledge of the types of data available, how the data were collected, and who collected it is very useful. It is recommended to have a dictionary that defines all data variables: it should contain the type of data, the range of expected values of each variable, general summary statistics, level of missingness, and subcomponents if available. The second recommendation is to develop a research plan that includes strategies for data quality appraisal and management such as statistical procedures for handling missing data and potential actions if other data quality issues arise (e.g., removal of extreme values, diagnostic code validation). The last recommendation is to promote transparency in reporting data quality including the proportion and type of missing data, other quality limitations, and any subsequent changes made to data values (e.g., variables removed for analysis, imputation methods, variable transformations, creation of new variables). This should enable the reuse of quality data for clinical research. Communications and sharing of the importance of data quality with clinicians are encouraged [28].

3 Clinical Coding Systems

In this section, we discuss clinical coding systems, classifications, or terminologies. We first introduce clinical coding systems and explain the motivation behind their existence and usage. This is followed by a discussion of the common attributes that coding systems tend to have, and how this relates to their usage for data analysis. We provide summaries for some of the most commonly used systems in use at the time of writing. Finally, we discuss some of the potential challenges and limitations of clinical coding systems.

3.1 Motivation

Recording clinical data using free text and local terminology incurs major barriers to conducting effective data analysis for health research [43]. Clinical coding systems significantly alleviate this problem, and so are of great usefulness to researchers and analysts when carrying out such work. Medical concepts are naturally described by linguistic terminology and are often associated with a descriptive text. Linguistic data is however loosely structured, and the same underlying medical concept might be expressed differently by different healthcare professionals. Clinical concepts can usually be expressed in a multitude of ways, both due to synonyms in individual terms and simply through different ways of combining and arranging words into a description. Processing large amounts of such data in order to perform modern computer-assisted data analysis, such as training machine learning models, would therefore require the use of natural language processing (NLP) techniques

[44]. Furthermore, when considering medical data from many countries, one would need to consider all the possible languages that medical records might be written in.

Instead of mapping clinical concepts into the highly complex realm of natural language, clinical coding systems seek to provide an unambiguous mapping from a given clinical concept to a unique encoding in a principled fashion. This makes it significantly easier to employ modern large-scale data analysis techniques on clinical data. For example, if one were interested in studying the prevalence of chronic fatigue, instead of having to attempt to exhaustively match records containing every conceivable way to express this linguistically, one would only need to identify which clinical codes are associated with the relevant clinical concepts and select records containing those codes.

3.2 Common Characteristics

Clinical coding systems can vary significantly in their descriptive scope, depending on their intended usage. The DSM-5 [45], for instance, limits its scope entirely to psychiatric diagnoses, while SNOMED-CT [46, 47] seeks to be as comprehensive as possible, including concept codes relating to, for example, body structure, physical objects, and environment. Both of these coding schemes describe concepts relevant at the level of individual patients, though codes can exist for broader or more fine-grained scopes such as public health or microbiology.

Typically, clinical coding schemes are arranged hierarchically, as this reflects the categorical relationship between clinical concepts well while also providing an intuitive means to find relevant concepts. This hierarchical structuring can be reflected in the identifiers used to encode clinical concepts, further aiding in their comprehension. In the ICD scheme [48], for example, codes begin with a character that identifies the relevant chapter in the ICD manual, and subsequent characters provide identification of finer and finer degrees of specification.

Another property of clinical coding systems that can be useful to classify is whether it is *compositional* or *enumerative* [49, Chapter 22]. In a compositional scheme, concepts can be encoded by combining more basic conceptual units together. This reduces the burden to specify large enough lists of distinct concepts to comprehensively cover all necessary clinical concepts required by scheme designers. This is in contrast to enumerative systems, which instead aim to achieve completeness by having a unique identifier for every concept within the scope of the scheme.

Clinical coding schemes can encode many kinds of relationships between concepts that are more specific than the simple parent-child relationship in basic hierarchies. These reflect the more nuanced kinds of relationships present in clinical concepts. Coiera [49, Chapter 22] outlines three main kinds of conceptual relationships: Part-Whole, Is-A, and Causal. Part-Whole

relationships are useful when a concept contains constituent parts which are also concepts, e.g., the eyes are a part of the face which is a part of the head which is a part of the body. This relationship is generally most useful for describing physical assemblages. Is-A relationships are perhaps the most common and indicate basic categorical similarities, such as Arterial Blood Specimen Is-A Blood Specimen Is-A Specimen. Finally, Causal relationships are used to indicate events or effects that arise as the result of another, or that cause another.

Hierarchical schemes may also introduce multiple axes upon which to expand concepts (essentially multiple hierarchies). In this way, elements belonging to a particular place in the hierarchy of one axis may also appear in the hierarchy of a different axis. This often involves a concept having multiple relationships of different types to a number of different concepts, i.e., a concept may have an Is-A relationship and a Causal relationship with two different concepts.

These are all useful features in the context of data science. Hierarchical structures allow for users of data to select as coarse or as fine-grained concepts as are relevant to their specific analyses. The defined relationships between concepts can be exploited in order to identify groups of relevant codes. Furthermore, some coding schemes, such as SNOMED-CT, may encode useful concepts beyond clinical events or concepts, such as whether patients have consented for research data usage, which can be useful, for example, in screening population members who are unsuitable for research cohorts, etc.

3.3 Notable Coding Systems

Here we provide summaries of commonly used coding systems that are likely to be encountered when performing analysis on EHR data. However, this is by no-means an exhaustive list. Many more are in use, and some datasets or corpora might use their own coding systems. In these cases, the data provider will usually specify mappings to more common systems such as ICD or SNOMED-CT. For example, in the case of the Clinical Practice Research Datalink (CPRD) [50], unique codes are provided for medical terms with mappings to Read Codes (a now largely legacy coding system in the United Kingdom), and unique treatment codes with links to the NHS Dictionary of Medicines and Devices (dm+d) [51] and the British National Formulary (BNF) [52], which provide codes relating specifically to medical products and prescribing.

3.3.1 SNOMED-CT

SNOMED-CT (Systematized Nomenclature of MEDicine-Clinical Terms) [46], maintained by SNOMED International, is a clinical coding scheme designed to be highly comprehensive and computer-processable. It is in wide usage around the world, in particular in the United Kingdom. SNOMED-CT supersedes the older SNOMED and SNOMED-RT systems. It is a hierarchical, compositional coding scheme, including specified relationships

Table 1
The top-level hierarchical categories in the SNOMED-CT system

Hierarchy
Body structure
Clinical finding
Event
Observable entity
Organism
Pharmaceutical/biologic product
Physical object
Procedure
Qualifier value
Situation with explicit context
Social context
Substance

between related concepts. It provides good linkage with ICD to allow for easy data sharing. There are 15 primary hierarchical categories in SNOMED-CT, to which all other concepts belong. A concept in SNOMED-CT is comprised of several elements. The primary identifying element is the Concept ID, which is a unique numerical identifier for the clinical concept. This is accompanied by a textual description of the concept. There are specified Relationships to other related concepts, and Reference Sets which provide groupings of concepts. SNOMED-CT codes are hierarchical and linked via *Is-A* relationships. Table 1 presents the top-level concepts of SNOMED-CT.

3.3.2 ICD

The ICD (International Classification of Diseases) [48] is a coding system created by the World Health Organization (WHO). While the ICD is currently in its 11th revision (ICD-11) [53], ICD-10 is still more commonly used at the time of writing, and the widespread adoption of ICD-11 will likely take more time. The ICD system is a multi-axis hierarchical coding system, assigning an alphanumeric code to each concept. Each code is procedurally derived from its concept's location in the hierarchy, aiding in comprehension. The first character letter in an ICD code associates it with a specific chapter in the ICD manual (see Table 2 for the different chapters of ICD-10). The following three characters locate the concept within the chapter and range from A00 to Z99. For more detail, each category can be further subdivided

Table 2
The chapters of ICD-10

Number	Chapter name
I	Certain infectious and parasitic diseases
II	Neoplasms
III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	Endocrine, nutritional, and metabolic diseases
V	Mental and behavioral disorders
VI	Diseases of the nervous system
VII	Diseases of the eye and adnexa
VIII	Diseases of the ear and mastoid process
IX	Diseases of the circulatory system
X	Diseases of the respiratory system
XI	Diseases of the digestive system
XII	Diseases of the skin and subcutaneous tissue
XIII	Diseases of the musculoskeletal system and connective tissue
XIV	Diseases of the genitourinary system
XV	Pregnancy, childbirth, and the puerperium
XVI	Certain conditions originating in the perinatal period
XVII	Congenital malformations, deformations, and chromosomal abnormalities
XVIII	Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified
XIX	Injury, poisoning, and certain other consequences of external causes
XX	External causes of morbidity and mortality
XXI	Factors influencing health status and contact with health services
XXII	Codes for special purposes

with up to three additional numeric characters. Table 3 shows multiple sclerosis as it appears in ICD-11 as an example of this hierarchical coding structure. The ICD system is intended to be limited in scope to disease diagnosis-related concepts; however, the WHO maintains additional systems to cover concepts outside of this scope. The ICF (International Classification of Functioning, Disability and Health), for instance, focuses on a patient's capacity to live and function and includes concepts relating to body functions, bodily structures, activities, participation, and environmental factors. Furthermore, various modifications of the ICD system exist to expand upon its capabilities for use in clinical settings, such as the ICD-10-CM in the United States and the ICD-10-CA in Canada.

Table 3
The hierarchical structure of multiple sclerosis within the ICD-11

1. ICD-11 for Mortality and Morbidity Statistics
• 08 - Diseases of the nervous system
– Multiple sclerosis or other white matter disorders
* 8A40 - Multiple sclerosis
· 8A40.0 - Relapsing-remitting multiple sclerosis
· 8A40.1 - Primary progressive multiple sclerosis
· 8A40.2 - Secondary progressive multiple sclerosis
· 8A40.Y - Other specified multiple sclerosis
· 8A40.Z - Multiple sclerosis, unspecified

3.3.3 UMLS

“The Unified Medical Language System (UMLS) is something like the Rosetta Stone of international terminologies”—Cocira [49, Chapter 23]

The UMLS [54] is intended to provide a means to relate coding systems to each other. It achieves this with three knowledge sources: the Metathesaurus, a semantic network, and the SPECIALIST Lexicon. The Metathesaurus is a nonhierarchical controlled vocabulary of terms organized by concept and provides the synonyms of concepts in different coding systems and is the primary way in which translation between systems is supported. Controlled vocabularies from hundreds of coding systems are represented in the Metathesaurus, and its entries are regularly updated. A complete list of all the supported controlled vocabularies is available in the UMLS Metathesaurus Vocabulary Documentation on the official website.¹ The Metathesaurus specifies defining attributes of concepts, and relationships between concepts, including Is-A, Part-Whole, and Causal relationship types. The semantic network provides the semantic types and relationships that concepts are permitted to inherit from. The primary semantic relationship is the hierarchical Is-A relationship, although there are five primary nonhierarchical relationship types: “physically related to,” “spatially related to,” “temporally related to,” “functionally related to,” and “conceptually related to.” The SPECIALIST Lexicon is intended to assist computer applications in interpreting free-text fields. It encodes syntactic, morphological, and orthographic information, including common spelling variants. In practice, most users of the UMLS do so indirectly through

¹ <https://www.nlm.nih.gov/research/umls/index.html>.

tools that rely on the UMLS, such as PubMed² and other clinical software systems such as EHR software and analysis pipelines. The most common uses are for extracting clinical terminologies from text and translating between coding systems [55].

3.3.4 Read Codes

Read Codes [56, 57] were used exclusively by the United Kingdom until 2018, when they were replaced by SNOMED-CT. Read Codes are organized hierarchically; however, the identifiers themselves do not indicate where in a hierarchy the concept belongs as they do in ICD. Version 3 (CTV-3) is the most recent version of Read Codes, and introduced compositionality to the system, while becoming less strictly hierarchical. Read Codes were intended to provide digital operability in primary care settings, but are no longer used in primary care in England (though they are still in use in Scotland at the time of writing and may be used in secondary care in England). Read Codes map well to ICD concepts. The Read Codes Drug and Appliance Dictionary is an extension of the Read Codes system to include pharmacological products, foods, and medical appliances for use in EHR software and prescribing systems.

3.4 Challenges and Limitations

The usefulness of clinical coding schemes is dependent upon their usage by healthcare professionals being thorough and appropriate. Improper usage of coding systems can occur, contributing to data quality issues such as incompleteness, inconsistency, and inaccuracy [58]. Further challenges can arise for researchers where data may contain multiple coding systems; this can happen if the data is collected from multiple different sources where different coding systems are in use, or if the period of data collection covers a change in the preferred coding system, such as the change from CTV-3 to SNOMED-CT in the United Kingdom. In these cases, the researcher must ensure that they consider relevant concepts from each different scheme or implement a mapping from one scheme to another. Most coding schemes provide good mapping support to ICD codes, and the UMLS coding system is designed to provide a means of translating between different schemes. Additionally, some sources of data may provide their own coding schemes that are not in usage (and thus not documented) elsewhere.

4 Protection and Governance of EHR Data

In this section, we will explore the focal points of data protection and governance analyzing the most recent jurisdictional background and its implication in real-world healthcare applications. In Subheading 4.1, we introduce the main legislative body and its

²<https://pubmed.ncbi.nlm.nih.gov/>.

core definitions in data protection. Then, Subheading 4.2 describes in a more technical way how data analysis can be conducted in a privacy-preserving manner.

4.1 Data Protection in a Nutshell

The explosive evolution of digital technologies and our ability to collect, store, and elaborate data is dramatically changing how we should consider privacy and data protection; particularly, the advent of artificial intelligence (AI) and advanced mathematical modeling tools made it necessary to reform the national and international data protection and governance rules to better protect people who generated such data and give them more control on what can be done with it. Although it is worth mentioning valuable independent contributions to the healthcare data protection guidelines like the Goldacre Review [59, 60], we will focus mainly on the most recent and structured action published at international level in terms of data protection and governance, the European General Data Protection regulation, or GDPR [17].

The GDPR was published by the European Commission in 2016 to set the guidelines that all member states must apply in their national legislation in terms of data protection. Although its legal validity is limited to the members of the European Economic Area (EEA), its effects expanded also to European Union (EU) candidate countries and the United Kingdom which embraced the new GDPR regulation through the UK GDPR [18] and maintained it part living of the legislation even after renouncing to the EU membership. It is worth mentioning that the effects of GDPR are not limited to the data management and governance executed within the countries that embrace the regulation, but is strictly related to the persons to whom the data belong; this means that the GDPR guidelines must be followed by any entity worldwide when dealing with data belonging to individuals from countries where the GDPR applies. GDPR defines as *personal data* any single information that is relatable to a person; in Box 1 we enumerate the three main agents required in any endeavor involving personal data management.

To contextualize these concepts in an healthcare scenario, if a non-European controller (e.g., an Australian hospital) aims at collecting, storing, or elaborating healthcare data from an individual protected by the GDPR or equivalent legislation for an international multicenter clinical trial, they still must respect all dictations of GDPR on that data specifically.

The GDPR reads: *personal data processing should be designed to serve mankind and the right to the protection of such data is not an absolute right, but must be considered in relation to its function in society*. Let's then consider this from the two angles of data governance and operation, and its purpose in the AI era.

Box 1: Basic agents recognized by GDPR

Data subject Individual(s) to whom the personal data belongs.

Controller Individual(s) or institution(s) responsible for implementing appropriate technical and organizational measures to ensure and to be able to demonstrate that processing is performed in accordance with the GDPR.

Processor Individual(s) or institution(s) responsible for using, manipulating, and leveraging personal data for the goals defined by the controller and agreed upon by the data subject.

4.1.1 Governance and Operation

One of the main dictations of GDPR is that data should be as anonymized (or, de-identified) and minimal as possible for a given application. This means that the data controller shall specify in details which data will be needed and why and collect only this required data, possibly in an anonymous way. Moreover, the data should be stored as long as the application requires it but not longer unless authorized by the data subject. This process should minimize as much as possible the identifiability of individuals, especially in those cases in which the content of data carries very sensitive information like health status, religious faith, political affiliation, and similar. Indeed, one of the main reasons why the use of free-text clinical notes in natural language processing (NLP) applications carries additional complications is that information that could identify individuals are often expressed in a nonstructured way in text (e.g., a specific reference to a person's habits, rare diseases, physical aspect, etc.) [61]. A similar issue arises with imaging applications, where the content of the imaging medical examination could contain personal information of its owner (e.g., the name written on an X-ray printing).

With a closer focus to EHR in a common tabular structure, identification of individuals can go beyond their names and unique identifiers. If the combination of other information can lead to their identification (e.g., the address, the sex, physical characteristics, profession, etc.), then the EHR is not technically anonymized. A step forward is the pseudo-anonymization, a process where the identifiable information fields are replaced with artificially created alternatives that encode or encrypt these information without direct disclosure. It is important to note that albeit this approach is valid in healthcare applications, it still allows a post hoc reconstruction of the identifiable data and should be implemented

carefully. Note that, in the specific case of brain images, the medical image may in principle allow reidentification of the patient (for instance, mainly through recognition of facial features such as the nose). For this reason, “defacing” (a procedure that modifies the image to remove facial features while preserving the content of the brain) is increasingly used. According to the Health Insurance Portability and Accountability Act of 1996 (HIPAA)³ issued by the US Department of Health and Human Services, 18 elements have to be deleted for an electronic health record to be considered de-identified; these include names, geographic subdivision smaller than a State, all elements of dates (except year) for dates directly related to an individual, telephone numbers, social security numbers, and license numbers. This practice can be exported internationally and used as a rule of thumb to ensure appropriate anonymization in all healthcare-related applications.

With respect to the many stages that comprise the analysis and elaboration of healthcare data, data protection can be handled in different and more flexible ways. Assuming a high level of internal protection of healthcare institutions (e.g., firewalls and encrypted servers), as long as the data remains within the institution secured information system, the majority of threats can be blocked and mitigated at an institutional level. Examples of threats are malicious access to and modification of data with the objective of compromising individual’s health or disrupting the operation of the hospital itself. The main exposure happens in case the data need to be transferred to another institution to carry out the required analyses. In this rather common case, the anonymization (or pseudo-anonymization) process should be carefully applied and data should never reside in a non-secured storage device or communication channel. To prevent this exposure to happen but at the same time to leave the possibility of leveraging the collected data for the purpose of AI applications and statistical studies, the federated learning methodology has been developed in recent years. This will be described further in Subheading 4.2.

The data subject has the right to get its own data deleted from the controller when, for example, the accuracy of the data is contested by the data subject, or when the controller no longer needs the data for its purposes. Similarly, the data subject has also the right to receive their personal data from the controller in a commonly used and machine-readable format and have then the right to transfer such data to another controller, when technically feasible, in a direct way. These aspects introduce operational constraints in EHR management as they require to be stored in an identifiable way (so as to allow its post hoc management, deletion, or

³ <https://www.hhs.gov/hipaa/index.html>.

modification) but to be elaborated in a non-identifiable manner to ensure that at any point of the data elaboration, the identification of the patients is impossible or as minimal as required for the elaboration itself. A corner case would be when a patient revokes the right of the controller to handle their data and its anonymized version is in use; from an operational point of view, this could cause the need for re-execution of the data extraction and elaboration.

4.1.2 The Purpose of EHR in the Era of AI

The main conundrum here is whether a specific use of healthcare data is functional to a societal benefit, which is a very difficult problem given its highly subjective interpretation. Indeed, as we continue producing beneficial applications, the opportunities to develop malevolent ones increase. Hostile actors may use private healthcare data and AI for personal profits, policy control, and other malicious cases. The availability of new tools suddenly sheds light on problems we didn't know we had and this is happening with AI and its application to healthcare. Machine learning and deep learning are by far the most successful technologies that are changing how we conceive data value and the importance of its quality [62], and when it comes to these computing tools, the more data, the better, but not only that; for each application, the data collected and elaborated should be as representative as possible of the learning task, which is a rather challenging issue considering the amount of human intervention in clinical data collection (especially in free-text annotations) and inherent biases in the data distribution over the available population. Current regulations are imposed to the data controllers to clearly communicate and have the explicit agreement of the data subject for any use they may do with it, and this is a fundamental protection of each individual's right to choose when and where their data can be used. This becomes particularly stressed in healthcare scenarios where misuse and abuses of patients' data can result in unethical advantage and/or enrichment of the institutions or individuals capable of making the most out of such abundant data.

Ethical approvals for the use of clinical datasets are usually granted by the hospitals' ethic committee, through detailed processes that every study has to undertake in its design phase. However, with increased focus on the use of AI technologies in medicine, the challenge becomes to contextualise within these ethics frameworks new technologies, the potential they carry, and the risks they may represent. Therefore, an integrated approach is needed between clinical experts, and AI/ML specialists to give more transparency, cohesion, and consistency to the use of data in health research.

4.2 Privacy-Preserving EHR Data Analysis

The training of any kind of AI-based predictive model requires as much data as possible, and given the nature of clinical data (costly and with high human intervention), it is often the case that a single healthcare institution is not enough to produce the data needed for the creation of a predictive model. This is particularly true in those cases in which the distribution of the population of patients within the hospital is not representative of the general population or at least of the possible population of patients for which those predictive models will be used.

The most straightforward practice to overcome this limitation consists in gathering data from multiple institutions in one single center and pre-process the data so as to integrate everything in one single training dataset. This allows the unification of the contribution of all healthcare institutions and therefore a more comprehensive, heterogeneous, and representative training dataset. Transferring clinical data from one hospital to another is a procedure that brings many privacy- and security-related problems, including the proper anonymization, or pseudo-anonymization, of clinical records and the encryption of the data en route to another institution.

The technical difficulties here dominate over the potential of a scalable, efficient, and secure data science pipeline that properly uses EHR to extract new knowledge and train predictive models.

One of the most brilliant solutions to solve these problems was initially proposed by Google with the federated learning methodology [63]. According to this approach designed primarily for deep neural networks, instead of transferring the data between institutions and collect everything in one unique dataset, a more efficient choice is to send the models to be trained to every institution that participates in the federation and, once one or more training steps are executed, gather the trained models in one central computing node (which can be one of the institutions) and compile the trained models in one comprehensive unique solution that represents the common knowledge produced.

Federated learning was designed for a task very different from clinical applications, i.e., the automatic completion of smartphones' keyboard, but its principles can be translated to the healthcare environment very effectively. The main benefits are that clinical data will never leave the owner's secured information system and anonymization and encryption of the data itself are not major problems. Moreover, the ability to involve the contribution of multiple centers for one training process requires a software infrastructure that can be utilized many more times for learning tasks.

4.3 Challenges Ahead

In the context of federated learning for EHR analysis, we find many challenges to be addressed in terms of both data quality and governance and learning methodologies. Here are listed some of the most relevant:

1. Not having direct access to other institutions' data makes it harder to assess the quality, consistency, and completeness of the datasets. This mandates additional care to the learning strategies as the representativeness of data must be preserved and phenomena like the *catastrophic forgetting* [64] produced by a large amount of data should be prevented.
2. Even assuming a good enough data quality in terms of completeness, correctness, and standards used, the distribution of data in independent datasets can be very different, posing additional learning challenges in the creation of a reliable and fair predictive model. This phenomenon is also known as the *non-IID*, or *non-independent and identically distributed*, data, and it is a very active research field [65].
3. Regardless of the immobility of data in healthcare information systems, the predictive models still have to travel between institutions, and this allows the possibility of data reconstruction through inverse gradient strategies [66], and the predictive model alteration (or *poisoning*) [67, 68] to induce it to behave in a malicious way; this transfers the security problems from the data to the machine learning models themselves and must be properly dealt both at a network level (with encrypted connections) and at a model level to mitigate communication bottlenecks, poisoning, backdoor, and inference-based attacks [69].

5 Conclusion

Increasing interest and opportunities for various research purposes were attracted by the rapidly growing number of EHRs. To draw valid and reliable research findings, data quality is paramount. In this chapter, we first introduced the definition of data quality, the reported components, and the concerns raised with poor data quality. Various aspects of data quality components and challenges were explored, such as data accuracy and data completeness. General practices for data quality analysis were recommended at the end of the data quality section.

We then introduced the concepts of a clinical coding system and discuss their potential challenges and limitations. We described the common characteristics of coding systems and then presented some of the most common ones: SNOMED-CT, ICD, UMLS, and Read Codes.

Finally, we navigated the main concepts of data governance and protection in healthcare settings. National and international regulations are put in place to define baseline principles to ensure the most appropriate treatment, storage, and final utilization of personal data, including healthcare information. From an operational perspective, there are numerous challenges to face, e.g., the

anonymization, or pseudo-anonymization, of patients' data and its proper privacy-preserving analysis for business and clinical purposes. This is particularly important in machine learning applications where a large amount of data is required and data sharing between hospitals is not a viable and secure solution. To produce a truly privacy-preserving approach for machine learning applications, federated learning is today the most effective and promising deployable methodology.

Acknowledgements

This research was funded/supported by the National Institute for Health and Care Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London and/or the NIHR Clinical Research Facility. WW and VC acknowledge the financial support from the Health Foundation. GHG and VC were supported by Innovation Scholars Big Data and AI Training MR/V038664/1 MRC funded by The Medical Research Council (MRC) at King's College London. The views expressed are those of the author(s) and not necessarily those of the Health Foundation, the MRC, the NHS, the NIHR, or the Department of Health and Social Care.

The authors are grateful to Prof. Paolo Missier for reviewing this chapter and providing useful insight.

References

1. CPRD (n.d.) Clinical practice research data-link. <https://cprd.com/>
2. QResearch (n.d.) QResearch. <https://www.qresearch.org/>
3. ResearchOne (n.d.) Transforming data into knowledge. <http://www.researchone.org/>
4. Alliance UHDR (2020) Hdruk innovation gateway — homepage. <https://www.healthdatagateway.org/>
5. Verheij R, van der Zee J (2018) Collecting information in general practice: “just by pressing a single button”? Morbidity, Performance and Quality in Primary Care pp 265–272. <https://doi.org/10.1201/9781315383248-36>
6. Nivel (n.d.) Nivel primary care database. <https://www.nivel.nl/en/nivel-zorgregistraties-eerste-lijn/nivel-primary-care-database>
7. Schweikardt C, Verheij RA, Donker GA, Coppieters Y (2016) The historical development of the dutch sentinel general practice network from a paper-based into a digital primary care monitoring system. *J Public Health (Germany)* 24:545–562. <https://doi.org/10.1007/S10389-016-0753-4/TABLES/3>. <https://link.springer.com/article/10.1007/s10389-016-0753-4>
8. Bartholomeeusen S, Kim CY, Mertens R, Faes C, Buntinx F (2005) The denominator in general practice, a new approach from the intego database. *Fam Pract* 22:442–447. <https://doi.org/10.1093/FAMPRA/CMI054>. <https://academic.oup.com/fampra/article/22/4/442/662730>
9. SNDS (n.d.) Système national des données de santé. <https://www.bordeauxpharmacoepi.eu/en/snds-presentation/>
10. Bezin J, Duong M, Lassalle R, Droz C, Pariente A, Blin P, Moore N (2017) The national healthcare system claims databases in France, SNIIRAM and EGB: powerful tools for pharmacoepidemiology. *Pharmacoepidemiol Drug Saf* 26(8):954–962
11. Daniel C, Salamanca E (2020) Hospital Databases: AP-HP data warehouse. In: Nordlinger B, Villani C, Rus D (eds)

- Healthcare and artificial intelligence. Springer, Berlin, pp 57–67
12. Ludvigsson JF, Almqvist C, Bonamy AKE, Ljung R, Michaësson K, Neovius M, Stephansson O, Ye W (2016) Registers of the Swedish total population and their use in medical research. *Eur J Epidemiol* 31(2):125–136
 13. Serda M (2013) Synteza i aktywność biologiczna nowych analogów tiosemikarbazonowych chelatorów żelaza
 14. Gliklich RE, Dreyer NA, Leavy MB (2014) Registries for evaluating patient outcomes. AHRQ Publication 1:669. <https://www.ncbi.nlm.nih.gov/books/NBK208616/>
 15. Fleurence RL, Beal AC, Sheridan SE, Johnson LB, Selby JV (2017) Patient-powered research networks aim to improve patient care and health research. *Health Aff* 33(7):1212–1219. <https://doi.org/10.1377/HLTHAFF.2014.0113>
 16. CTSA (n.d.) CTSA Central. <http://www.ctsacentral.org/>
 17. GDPR (2016) EU General Data Protection Regulation. <http://data.europa.eu/eli/reg/2016/679/oj>
 18. UK GDPR (2018) UK General Data Protection Regulation Updated for Brexit — UK GDPR. <https://uk-gdpr.org/>
 19. Fundation TM (2006) Background issues on data quality. In: The connecting for health common framework <https://bok.ahima.org/PdfView?oid=63654>
 20. Feder SL (2018) Data quality in electronic health records research: quality domains and assessment methods. *West J Nurs Res* 40(5):753–766. <https://doi.org/10.1177/0193945916689084>
 21. Chan KS, Fowles JB, Weiner JP (2010) Review: Electronic health records and the reliability and validity of quality measures: A review of the literature. *Med Care Res Rev* 67(5):503–527. <https://doi.org/10.1177/1077558709359007>
 22. Kahn M, Raebel M, Glanz J, Riedlinger K, Steiner J (2012) A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care* 50(Suppl):S21–9. <https://doi.org/10.1097/MLR.0b013e318257dd67>
 23. Wand Y, Wang RY (1996) Anchoring data quality dimensions in ontological foundations. *Commun ACM* 39(11):86–95. <https://doi.org/10.1145/240455.240479>
 24. Weiskopf NG, Weng C (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 20(1):144–151. <https://doi.org/10.1136/amiajnl-2011-000681>. <https://academic.oup.com/jamia/article-pdf/20/1/144/9517051/20-1-144.pdf>
 25. Ahmad F, Rasmussen L, Persell S, Richardson J, Liss D, Kenly P, Chung I, French D, Walunas T, Schriever A, Kho A (2019) Challenges to electronic clinical quality measurement using third-party platforms in primary care practices: The healthy hearts in the heartland experience. *JAMIA Open* 2(4):423–428. <https://doi.org/10.1093/jamiaopen/ooz038>
 26. Tse J, You W (2011) How accurate is the electronic health record?—a pilot study evaluating information accuracy in a primary care setting. *Stud Health Technol Inform* 168:158–64
 27. Ozair F, Nayer J, Sharma A, Aggarwal P (2015) Ethical issues in electronic health records: A general overview. *Perspect Clin Res* 6:73–6. <https://doi.org/10.4103/2229-3485.153997>
 28. Bayley K, Belnap T, Savitz L, Masica A, Shah N, Fleming N (2013) Challenges in using electronic health record data for CER: Experience of 4 learning organizations and solutions applied. *Med Care* 51:S80–S86. <https://doi.org/10.1097/MLR.0b013e31829b1d48>
 29. Hyun K (2013) The prevention and handling of the missing data. *Korean J Anesthesiol* 64(5):402–406. <https://doi.org/10.4097/kjae.2013.64.5.402>. <http://ekja.org/journal/view.php?number=7569>
 30. Rubin DB (1976) Inference and missing data. *Biometrika* 63(3):581–592. <http://www.jstor.org/stable/2335739>
 31. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 338. <https://doi.org/10.1136/bmj.b2393>. <https://www.bmjjournals.org/content/338/bmj.b2393>. <https://www.bmjjournals.com/content>
 32. Smith WG (2008) Does gender influence online survey participation? A record-linkage analysis of university faculty online survey response behavior. Online Submission
 33. Little R, Rubin D (2002) Statistical analysis with missing data. In: Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, London. <http://books.google.com/books?id=aYPwAAAAMAAJ>
 34. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via

- the em algorithm. *J R Stat Soc Ser B Methodol* 39(1):1–38. <http://www.jstor.org/stable/2984875>
35. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P (2013) Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale J Biol Med* 86(3):343–358. <https://europepmc.org/articles/PMC3767219>
 36. Jakobsen JC, Gluud C, Wetterslev J, Winkel P (2017) When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. *BMC Med Res Methodol* 17(1):1–10
 37. Zhang Y, Flórez ID, Lozano LEC, Aloweni FAB, Kennedy SA, Li A, Craigie SM, Zhang S, Agarwal A, Lopes LC, Devji T, Wiercioch W, Riva JJ, Wang M, Jin X, Fei Y, Alexander PE, Morgano GP, Zhang Y, Carrasco-Labra A, Kahale LA, Akl EA, Schüemann HJ, Thabane L, Guyatt GH (2017) A systematic survey on reporting and methods for handling missing participant data for continuous outcomes in randomized controlled trials. *J Clin Epidemiol* 88:57–66
 38. Jørgensen AW, Lundstrøm LH, Wetterslev J, Astrup A, Gøtzsche PC (2014) Comparison of results from different imputation techniques for missing data from an anti-obesity drug trial. *PLoS One* 9(11):1–7. <https://doi.org/10.1371/journal.pone.0111964>
 39. Sinharay S, Stern H, Russell D (2001) The use of multiple imputation for the analysis of missing data. *Psychol Methods* 6:317–29. <https://doi.org/10.1037/1082-989X.6.4.317>
 40. Azur M, Stuart E, Frangakis C, Leaf P (2011) Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 20:40–9. <https://doi.org/10.1002/mpr.329>
 41. Horton NJ, Lipsitz SR (2001) Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am Stat* 55(3):244–254. <http://www.jstor.org/stable/2685809>
 42. Little RJA, Wang Y (1996) Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* 52(1):98–111. <http://www.jstor.org/stable/2533148>
 43. Elkin PL, Trusko BE, Koppell R, Speroff T, Mohrer D, Sakji S, Gurewitz I, Tuttle M, Brown SH (2010) Secondary use of clinical data. *Stud Health Technol Inform* 155:14–29
 44. Koleck TA, Dreisbach C, Bourne PE, Bakken S (2019) Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc* 26(4):364–379. <https://doi.org/10.1093/jamia/ocy173>
 45. Association AP, Association AP (eds) (2013) Diagnostic and statistical manual of mental disorders: DSM-5, 5th edn. American Psychiatric Association, Arlington, VA, oCLC:830807378
 46. SNOMED International (2022) SNOMED CT. <https://www.nlm.nih.gov/healthit/snomedct/index.html>. publisher: U.S. National Library of Medicine
 47. Lee D, de Keizer N, Lau F, Cornet R (2014) Literature review of SNOMED CT use. *J Am Med Inform Assoc* 21(e1):e11–e19. <https://doi.org/10.1136/amiajnl-2013-001636>
 48. World Health Organisation (2022) International Classification of Diseases (ICD). <https://www.who.int/standards/classifications/classification-of-diseases>
 49. Coiera E (2015) Guide to health informatics. CRC Press, Boca Raton. google-Books-ID: IngZBwAAQBAJ
 50. Medicines and Healthcare products Regulatory Agency (2022) Clinical Practice Research Datalink | CPRD. <https://www.cprd.com>
 51. NHS (2022) Dictionary of medicines and devices (dm+d) — nhsbsa. <https://www.nhsbsa.nhs.uk/pharmacies-gp-practices-and-appliance-contractors/dictionary-medicines-and-devices-dmd>
 52. Committee JF (2022) BNF (British national formulary) — nice. <https://bnf.nice.org.uk/>
 53. Organisation WH (ed) (2019) International statistical classification of diseases and related health problems, 11th edn. World Health Organization, New York. <https://icd.who.int/>
 54. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 32 (Database issue):D267–D270. <https://doi.org/10.1093/nar/gkh061>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308795/>
 55. Amos L, Anderson D, Brody S, Ripple A, Humphreys BL (2020) UMLS users and uses: a current overview. *J Am Med Inform Assoc* 27(10):1606–1611. <https://doi.org/10.1093/jamia/ocaa084>
 56. NHS (2020) Read Codes. <https://digital.nhs.uk/services/terminology-and-classifications/read-codes>
 57. N B (1994) What are the Read Codes? *Health Libr Rev* 11(3):177–182. <https://doi.org/10.1046/j.1365-2532.1994.1130177.x>. <https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2532.1994.1130177.x>
 58. Botsis T, Hartvigsen G, Chen F, Weng C (2010) Secondary use of EHR: data quality

- issues and informatics opportunities. Summit on Translational Bioinformatics 2010:1–5. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041534/>
59. Ben Goldacre ea (2022a) Better, broader, safer: using health data for research and analysis—gov.uk. <https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis>
 60. Ben Goldacre ea (2022b) Home — goldacre review. <https://www.goldaccreview.org/>
 61. Pan X, Zhang M, Ji S, Yang M (2020) Privacy risks of general-purpose language models. Proceedings—IEEE Symposium on Security and Privacy 2020(May):1314–1331. <https://doi.org/10.1109/SP40000.2020.00095>
 62. Esteve A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J (2019) A guide to deep learning in healthcare. Nat Med 25(1):24–29. <https://doi.org/10.1038/s41591-018-0316-z>. <https://www.nature.com/articles/s41591-018-0316-z>
 63. McMahan B, Moore E, Ramage D, Hampson S, Arcas BAy (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Singh A, Zhu J (eds) Proceedings of the 20th international conference on artificial intelligence and statistics, PMLR, Proceedings of Machine Learning Research, vol 54, pp 1273–1282. <https://proceedings.mlr.press/v54/mcmahan17a.html>
 64. McCloskey M, Cohen NJ (1989) Catastrophic interference in connectionist networks: the sequential learning problem. In: Bower GH (ed) Psychology of learning and motivation, vol 24, Academic Press, New York, pp 109–165. [https://doi.org/10.1016/S0079-7421\(08\)60536-8](https://doi.org/10.1016/S0079-7421(08)60536-8). <https://www.sciencedirect.com/science/article/pii/S0079742108605368>
 65. Zhu H, Xu J, Liu S, Jin Y (2021) Federated learning on non-IID data: A survey. Neurocomputing 465:371–390. <https://doi.org/10.1016/j.neucom.2021.07.098>, 2106.06843
 66. Geiping J, Bauermeister H, Dröge H, Moeller M (2020) Inverting gradients—how easy is it to break privacy in federated learning? In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates Inc, New York, pp 16937–16947. <https://proceedings.neurips.cc/paper/2020/file/c4ede56bbd98819ae6112b20ac6bf145-Paper.pdf>
 67. Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V (2020) How to backdoor federated learning. In: Chiappa S, Calandra R (eds) Proceedings of the twenty third international conference on artificial intelligence and statistics, PMLR, proceedings of machine learning research, vol 108, pp 2938–2948. <https://proceedings.mlr.press/v108/bagdasaryan20a.html>
 68. Lyu L, Yu H, Zhao J, Yang Q (2020) Threats to federated learning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12500 LNCS:3–16. https://doi.org/10.1007/978-3-030-63076-8_1. <https://arxiv.org/abs/2003.02133v1>, 2003.02133
 69. Mothukuri V, Parizi RM, Pouriyeh S, Huang Y, Dehghantanha A, Srivastava G (2021) A survey on security and privacy of federated learning. Futur Gener Comput Syst 115:619–640. <https://doi.org/10.1016/j.future.2020.10.007>. <https://www.sciencedirect.com/science/article/pii/S0167739X20329848>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 12

Mobile Devices, Connected Objects, and Sensors

Sirenia Lizbeth Mondragón-González, Eric Burguière, and Karim N'diaye

Abstract

Brain disorders are a leading cause of global disability. With the increasing global proliferation of smart devices and connected objects, the use of these technologies applied to research and clinical trials for brain disorders has the potential to improve their understanding and create applications aimed at preventing, early diagnosing, monitoring, and creating tailored help for patients. This chapter provides an overview of the data these technologies offer, examples of how the same sensors are applied in different applications across different brain disorders, and the limitations and considerations that should be taken into account when designing a solution using smart devices, connected objects, and sensors.

Key words Smartphone, Mobile devices, Wearables, Connected objects, Brain disorders, Digital psychiatry, Digital neurology, Digital phenotyping, Machine learning, Human activity recognition

1 Introduction

Sensors are devices that detect events or significant changes in their environment and send the information to other electronic devices for signal processing. Since they surround us continuously, we have integrated them so naturally into our lives that we are mostly unaware of their continuous functioning. They exist in everyday objects, from the motion unit installed in your mobile phone that allows you to switch from landscape to portrait view by simply rotating it to the presence detector sensor in your building that switches the light on and off. Indeed, there is a good chance that you are using one or multiple sensors right now without noticing. They provide various means to measure characteristics related to a person's physiology or behavior either in a laboratory/healthcare unit or in their daily life. They have thus raised a major interest in medicine in the past years. They are particularly interesting in the context of brain disorders because they allow monitoring of clinically relevant characteristics such as movement, behavior, cogni-

tions, etc. This chapter provides an introduction to the use of sensors in the context of brain disorders. The remainder of this chapter is organized as follows.

Subheading 2 presents an overview of the various data types collected using mobile devices, connected objects, and sensors that are relevant to brain disorder research and related clinical applications, in particular for machine learning (ML) processing. The relevance of these ubiquitous sensors comes from the possibility of collecting large amounts of data, allowing the continuous documentation of the user's daily life, an often critical issue with ML applications. Subheading 3 describes how these technologies might serve such applications in brain disorder research and clinics. Because of the strategic importance of ML in the on-device experience, mobile manufacturers have recently started to design and include specially designed microprocessors for ML calculations in smartphones and tablets, benefiting the third-party app development community. A different approach consists of cloud offload processing allowing lighter wearables and handheld devices. The main public interest in current applications of ML is to help guess what is expected by the user, eliminating the number of actions and decisions we make each day (facial recognition for security instead of remembering a password, classification in your picture gallery according to names or faces, recommending songs to listen based on your history and ratings, etc.). Although decision support might not necessarily be its first goal, the scholar community interested in brain disorders must be familiarized with this ongoing ML revolution since the technology is already there, opening the way to unprecedented opportunities in research and clinics. Subheading 4 describes limitations, caveats, and challenges that researchers willing to use such technologies and data need to be aware of.

2 Data Available from Mobile Device, Sensors, and Connected Objects for Brain Disorders

Far from presenting an extensive list of available sensors and devices, we aim to introduce the type of data one can exploit and sketch possible applications relevant to brain disorder research. The kind of data that we present here comes from sensors that are typically used for human activity recognition (HAR) or that we deemed relevant for the scope of this book. In particular, we have purposely omitted connected technologies that are used by health practitioners or in healthcare units and that require medical or specific training for their use and interpretation and that are therefore not commonly available to the public, such as wireless electroencephalographer (EEG—but *see* Chapter 9 for in-depth coverage of ML applied to EEG). We also set aside mobile technologies that

are not directly aimed at probing brain and behavioral functions, such as blood pressure monitor devices, glucometers, etc. (*see [1]* for a review).

We present these data types in two groups according to the role that the user (e.g., the patient) takes in acquiring the data: active vs. passive. In this context, we mainly describe typical applications, but we also point the readers to specific applications for which active or passive data can be used. For instance, vocal recordings can be actively collected by instructing the user to self-record (e.g., when completing a survey), but a microphone may also passively and continuously record the sound environment without the user triggering it (e.g., automatic handwashing recognition using the microphone of the Apple watch to detect water sound [2]). To explore the possibilities in data collection, we distinguish three interconnected elements: the person of interest, the device (including its potential interface), and the environment. According to the dimension of interest, we can focus on the data obtained from the interaction between these three elements (*see Fig. 1*).

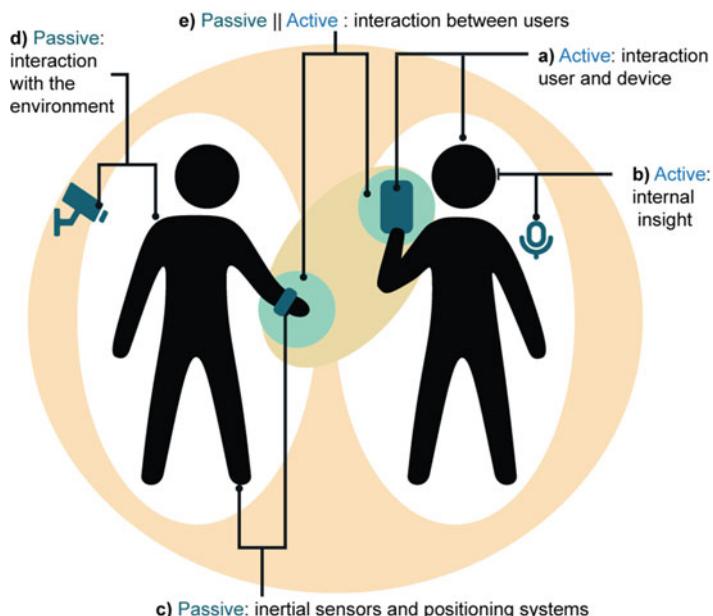


Fig. 1 *Active and passive sensing.* Mobile devices and wearable sensors provide metrics on various aspects of the mental and behavioral states through active (requiring an action from the user, often following a prompt) or passive (automatically without intentional action from the user) data collection. This is possible through (a) direct interaction with the device, (b) active use of a device for assessment of internal insight, (c) passive use of inertial and positioning systems, (d) passive interaction with sensors embedded in the environment, and (e) passive or active interaction between users through devices

2.1 Active Data Probing

In active data probing, the person of interest must execute a specific action to supply the data, meaning that the quantity and the quality of acquired data directly depend on the user's compliance. These actions usually involve direct interaction with the device. To maximize compliance, the subject needs to spend time and energy collecting the data; therefore, the number of action steps necessary to enter the data must be optimized to avoid user fatigue. It is also essential to care for feature overload by focusing on usability instead of utility and thoughtfully circumscribing the scope of questions or inputs. The amount of information requested and the response frequency are essential aspects to think ahead to maximize the continuous use of the device. If there is an intermediate user interface, following standard UX/UI (user experience/user interface) guidelines is a good starting point for optimization but might not be sufficient according to the target population group. It is crucial to design without making assumptions but by getting patients' early feedback through co-construction or participatory design [3–5]. In summary, there are several considerations that one needs to plan before deploying a solution-using active probing that involves the device itself but also how the user interacts with it.

2.1.1 Interaction with the User

Recording the subject's response can provide unique information about the occurrence of experiences and the cognitive processes that unfold over time. We can record the user's feedback at specific points in time or continuously by taking advantage of the interaction between the user and a device (see Fig. 1a).

Manual devices: response buttons, switches, and touchscreens. These devices capture conventional key or screen presses via switches or touchscreens, usually operated by hand. A switch connects or disconnects the conducting path in an electrical circuit, allowing the current to pass through contacts. They allow a subject to send a control or log signal to a system. They have been largely used, for several decades, in computer-based experiments for psychology, psychophysiology, behavioral, and functional magnetic resonance imaging (fMRI) research. The commonly obtained metrics are specific discrete on/off responses (pressed or not) and reaction time [6]. It is usually necessary to measure a person's reaction time to the nearest millisecond which requires dedicated response pads. Indeed, general-purpose commercial keyboards and mice have variable response delays ranging from 20 to 70 ms, a range comparable to or lower than human reaction time in a simple detection task [7]. On the other hand, dedicated computerized testing devices seek to have less variable and smaller response delay. They introduce less variation and biases in timing measurements [7] by addressing problems such as mechanical lags, debouncing, scanning, polling, and event handling. Commercially available response-button boxes (e.g., Psychology Software Tools, Inc., Sharpsburg, PA, USA; Cedrus Corporation, San Pedro, CA, USA; Empirisoft Corporation, New York, NY, USA; Engineering

Solutions, Inc., Hanover, MD, USA; PsyScope Button Box by New Micros in Dallas, TX, USA) have few options and specific layouts to collect responses according to standard gamepad layouts while still being usually customizable for more specific applications.

Alternatively, touchscreens can be used to detect discrete responses with screen coordinates of the touch or pressing. They come in many forms, and the most popular type works with capacitive or resistive sensors. Resistive touchscreens are pressure-sensitive, and capacitive screens are touch-sensitive. Nowadays, capacitive screens are more used because of their multi-touch capabilities, short response time, and better light transmission. However, if an application needs the exact coordinates of the contact, the inductive touchscreens are more suited. This technology is usually featured in the highest priced tablets along with a special pen that induces a signature electromagnetic perturbation that improves its precision compared to finger pointing. The disadvantage of touchscreens is that they lack tactile feedback and have high energy consumption. For collecting continuous responses, a joystick, computer mouse, or touchscreen may be used to track movement trajectories supposedly reflecting the dynamics of mental processes [8].

Connected devices have been introduced in many domains of everyday life and, more recently, in health and research settings, sometimes with medical-grade applications [9]. Such devices may include sensors of health-relevant physiological parameters (e.g., weight, heart rate, and blood pressure) or health-related behaviors (e.g., treatment compliance). These connected systems make data collection more systematic and readily available to the clinical practitioner. They are automatically integrated into data management systems. For example, on a pre-specified schedule, the patient will measure his/her blood pressure with a so-called smart blood pressure monitor, which may provide reminders and record and transmit these measurements to his/her doctor. Active connected devices (which require the patient to participate in the data collection process) may also track behavior: a connected pillbox would allow monitoring that the patient takes the medication according to the prescribed schedule [10]. In a subsequent part of this chapter, we will refer to passive connected medical devices (which perform measurements without the intervention of the user/patient), such as fall detection systems.

2.1.2 Subjective Assessments

With current knowledge and technologies, data that reflect psychological states such as emotions and thoughts can only be obtained by active data probing of the patient or an informer, usually a partner, family member, or caregiver (*see* Fig. 1b). The long history of psychological assessment provides rich conceptual and methodological frameworks for collecting valid measures of subjective

states when collected with a traditional semi-directed interview or paper-and-pencil questionnaires. Nevertheless, the novel possibilities allowed by mobile technologies challenge those traditional well-validated assessment tools by renewing the format and the content of questions addressed to the user. In medical care and research, patient-reported outcomes are at the heart of a paradigmatic change in medicine and clinical research, where patient-centric measures tend to be favored over pure biomedical targets.

Subjective assessments may sometime take the form of utterances or text. For machine learning applications, those have to be converted into data usable for feeding mathematical models. Natural language processing (NLP) tools have recently made substantial progress thanks to deep learning techniques, making even complex spontaneous oral or written language amenable to machine processing [11].

2.2 Passive Data Probing

In passive data probing, the data is collected without explicitly asking the subject to provide the data. It provides an objective representation of the subject's state in time. In scenarios where the data needs to be acquired multiple times a day, passively collecting the data is a more valuable and ecological way to proceed. It allows objectively measuring the duration and frequency of specific events and their evolution in time. In contrast with active data, probing can provide more samples over a period. Since meaningful events might be embedded in the collected data, this probing type requires reviewing historical loggings or computer applications to extract the information of interest.

2.2.1 Inertial and Positioning Systems

Detection of whole-body activities (such as walking, running, and bicycling), as well as fine-grained hand activity (such as smartphone scrolling, typing, and handwashing), can allow the arduous task of studying and monitoring human behavior, which is of great value to understand, prevent, and diagnose brain diseases as well as to provide care and support to the patient. The change in physical activity and its intensity, the detection of sleep disorders, fall detection, and the evolution or detection of a particular behavior are some possibilities that can be assessed with inertial sensors.

Identifying specific activities of a person based on sensor data is the main focus of the broad field of study called human activity recognition (HAR). A widely adapted vehicle for achieving HAR's goal is passive sensor-based systems that use inertial sensors (*see* Fig. 1c), which transduce inertial force into electrical signals to measure the acceleration, inclination, and vibration of a subject or object (*see* Fig. 2a). These systems are commonly included in today's portable electronic devices such as mobile phones, smartwatches, videogame controllers, clothes, cameras, and non-portable objects like cars and furniture. Besides offering the advantage, due to their reduced size, of being embeddable in

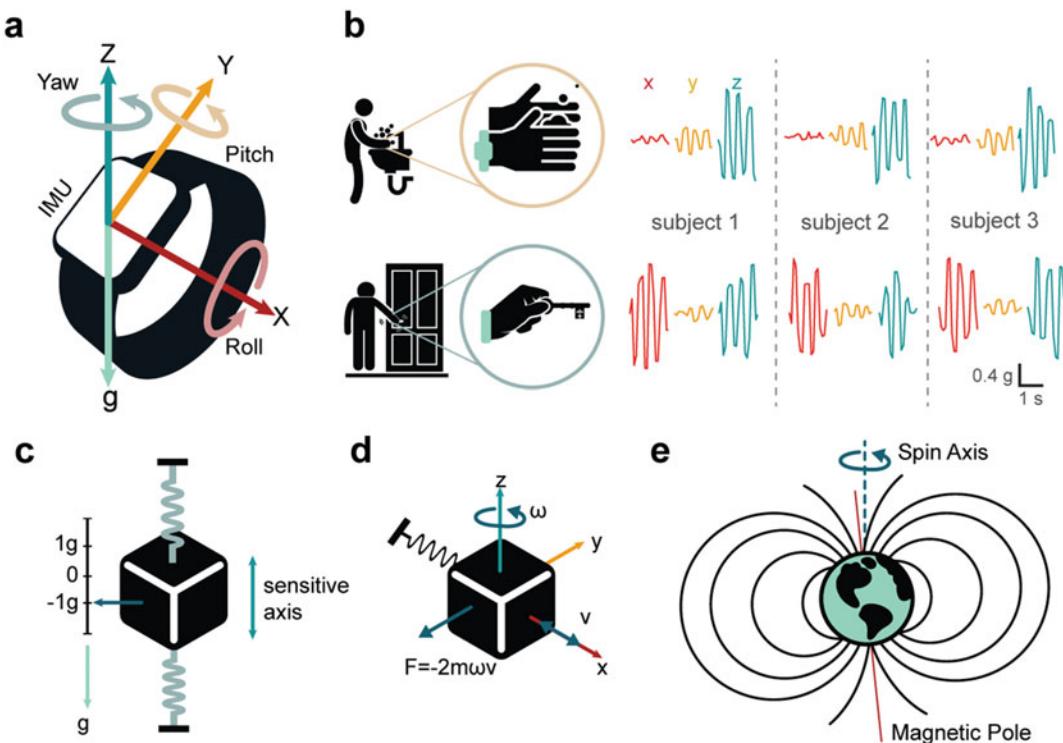


Fig. 2 Inertial sensors. (a) Representation of an inertial measurement unit (IMU) depicting the sensing axes and the corresponding yaw, pitch, and roll rotations. (b) Exemplar accelerometer profiles of two hand gestures (hand rubbing and key locking) for three subjects showing the similar periodic nature of the hand movements. (c) Operating principle of an MEM accelerometer. When a force is detected due to a compressive or extensive movement, it is possible to determine the displacement x and acceleration since the mass and spring constants are known. (d) Representation of a simple gyroscope model. (e) The magnetic field generated by electric currents, magnetic materials, and the Earth's magnetic force exerts a magnetic force detectable by a magnetometer sensor

almost any possible device, they are perceived as less intrusive of personal space than other HAR systems, such as camera and microphone-based systems [12], allowing to sense more naturalistic motion information uninterrupted. Most prior work on activity detection has focused on detecting whole-body activities that reflect ambulatory states and their degree of locomotion or lack of it, such as running, walking, cycling, lying, climbing stairs, falling, sitting, standing, and monitoring the sleep-wake cycle. Whole-body activities differ from fine-grained human actions, usually undertaken by the hands (*see* Fig. 2b). These hand activities are often independent of whole-body activity, for instance, sending a text from your smartphone while walking. A sustained sequence of related hand gestures composes a hand activity. Hand gestures like waves, flicks, and snaps tend to have exaggerated motions (used for communications), and hand activities are more subtle, discontinuous, and of varying durations [12]. Examples of complex hand

gestures are writing, typing, painting, searching the Internet, smoking, eating, and drinking. The way one approaches whole-body activity detection differs from fine-grained activity recognition in terms of the analysis approach (e.g., selected features), sensor configuration (e.g., higher sampling frequency for fine-grained activities than for whole-body activities), and location on the body (e.g., wrist vs. hip). In both detection problems, the most common sensors used for HAR applications are accelerometers, gyroscopes, and magnetic sensitive sensors (*see* Fig. 2c–e).

Accelerometers

Accelerometers are sensors used to measure linear acceleration, viz., change in velocity or speed per time interval of the object being measured along reference axes. Furthermore, one can obtain velocity information by integrating accelerometry data with respect to time. The measuring acceleration unit in the International System of Units (SI) is a meter per second squared (m/s^2). Since we can distinguish a static component in the accelerometer signal as the gravitational acceleration, it is also common to use the unit G-force (g) to distinguish the relative free-fall gravitational acceleration with a conventional standard value of $1\text{ g} = 9.81\text{ m/s}^2$. A simplistic representation of the accelerometer's operation principle is based on a suspended mass attached to a mechanical suspension system with respect to a reference inside a box, as shown in Fig. 2c. The inertial force due to gravity or acceleration will cause the suspended mass to deflect according to Hooke's law ($F = mk$) and Newton's second law ($F = ma$), where F denotes the force (N), m is the mass of the system (kg), k is the spring constant, x is the displacement (m), and a is the acceleration (m/s^2). This acceleration force can then be measured electrically with the changes in mass displacement with respect to the reference. To better understand this working principle, you can think of your experience as a passenger in a car rapidly moving back and forth and how the forces acting on you make you incline backward and forward on your seat. In nowadays-electronic devices, we find mostly miniaturized semiconductor accelerometers (microelectromechanical systems or MEMs), which are small mechanical and electrical devices mounted on a silicon chip. The most common types are piezoresistive, piezoelectric, and differential capacitive accelerometers [13]. Since the accelerometer is usually a built-in component embedded in a mobile device, the data we can obtain is provided in the XYZ coordinate system of the accelerometer component. The XYZ orientation is specific to each device, and its coordinate system is found in the datasheets of the components.

When processing the accelerometer signals, separating the acceleration due to movement from gravitational acceleration and noise sources (e.g., electronic device and measurement conditions) is necessary. A low-pass filter with a cutoff frequency of 0.25–3 Hz is usually applied to raw data to remove noise [14]. Alternatively,

transforming the raw accelerometer data to the vector magnitude (*Eq. 1*), which measures the instantaneous intensity of the subject's movement at time t can be done before filtering to remove noise and/or gravity from body acceleration. The following processing steps usually include normalization (min–max, division by maximum absolute value, or division by the mean).

$$vm(t) = \sqrt{A_x(t)^2 + A_y(t)^2 + A_z(t)^2} \quad (1)$$

A time-window segmentation is often necessary to retrieve information from the accelerometer time series. The epochs are usually consecutive sliding windows with an overlapping percentage (usually 50% overlap). Different window sizes can be compared to identify the optimal size for HAR analysis.

Gyroscope

A gyroscope is an inertial sensor that measures the rate of change of the angular position over time with respect to an inertial reference frame, also known as angular velocity or angular rate. The principle of function of MEM's gyroscopes is based on the Coriolis effect, which acts on moving objects within a frame of reference that rotates with respect to an inertial frame. Figure 2d represents a simple gyroscope model where a mass suspended on springs has a driving force on the x -axis and angular velocity ω applied about the z -axis, causing the mass to experience a force in the y -axis as a result of the Coriolis force. In an MEM's gyroscope, the resulting displacement is measured by a capacitive sensing structure. The angular velocity unit is deg./s, but expressing it in radians per second (rad/s) is also common. A gyroscope can provide information about activities that involve rotation around a particular axis. A triaxial gyroscope can provide information from three different angles, pitch (x -axis), roll (y -axis), and yaw (z -axis), to help estimate the movement signature's orientation and rotation.

In human activity recognition, the gyroscope activity helps provide information about activities involving rotation around a particular axis. While a gyroscope has no initial frame of reference like gravity, it can be combined with accelerometer data to measure angular position and help determine an object's orientation within 3D space. To obtain the angular position, we can integrate the angular velocity with Eq. 2, where p = yaw, pitch, and roll and θ_{p_0} is the initial angle compared to the Earth's axis coordinates.

$$\theta(t) = \int_0^t \dot{\theta}_p(t) dt + \theta_{p_0} \quad (2)$$

When the changes in angular velocity are faster than the sampling frequency, one will not be able to detect them, and the error will continue to increase with time. This error is called drift. Therefore, the sampling rate value should be carefully chosen since gyroscopes are vulnerable to drifting over the long term.

Magnetic Sensitive Sensors (e.g., Hall Sensor)

Magnetic sensors measure the strength and direction of the Earth's magnetic field and are affected by electric currents and magnetic materials (see Fig. 2e). Most MEM's magnetic sensors are based on magnetoresistance to measure the surrounding magnetic field, meaning that the resistance changes due to changes in the Earth's and nearby magnetic fields. They can detect the vector characterized by strength and direction toward the Earth's magnetic north, and with it, one can estimate one's heading. This vector is vertical at the Earth's magnetic pole and has an inclination angle of 0°. When used with accelerometers and gyroscopes, it can help to determine the absolute heading.

IMU Technology

The combination of accelerometers, gyroscopes, and sometimes magnetometers in a single electronic device is referred to as an inertial measuring unit (IMU). Here are some considerations when choosing an IMU system or a device that contains an accelerometer, gyroscope, or magnetic sensor for HAR applications:

1. *Dynamic range.* Dynamic range refers to the range of maximum amplitude that the sensor can measure before distortion. In the case of accelerometers, where the amplitude in locomotion increases in magnitude from cranial toward caudal body parts, they are typically measured in powers of two ($\pm 2\text{G}$, $\pm 4\text{G}$, $\pm 8\text{G}$, and so on), with an amplitude range of $\pm 12\text{G}$ for whole-body activities [15]. Gyroscopes are grouped by the angular rotation rate they can quantify (in thousands of degrees/second). The measuring range of magnetometers is in milliTesla (mT).
2. *The number of sensitive axes.* Inertial units that can sense in three orthogonal planes (triaxial) are suitable for most applications since different directions contribute to the total complex movement patterns.
3. *Bandwidth.* The sampling rate determines the frequency range that can be represented in a waveform. Its unit is samples per second or Hertz. For HAR applications, the bandwidth of human accelerations of interest must be covered by the sensor's sampling rate. The sampling rate selection depends on the activity of interest, the measured axes, and the body part to which the sensor is attached. For instance, walking at natural velocity ranges from 0.8 to 5 Hz when measured in the upper body, whereas abrupt accelerations up to 60 Hz have been measured at the foot level [15]. For typical whole-body activities (like lying, sitting, standing, and walking), sampling rates are usually between 50 and 200 Hz. Still, some studies use low ranges 20–40 Hz or as high as 4 kHz [12, 16] with analysis window lengths from 2 to 15 s [14]. A study has reported that frequencies from 0 to 128 Hz best characterize most human activities via hand monitoring [12].

4. *Interface and openness.* In HAR applications, IMUs interface with other systems for signal processing. It is essential to know the communication protocol for data transfer and the degree of openness of the chosen system to allow configuration and extraction of raw signals since not all commercial systems allow raw signal extraction or changes in some parameters, like the sampling frequency.
5. *Sensor biases.* Sensor bias refers to the initial offset in the signal output when there is no movement. In the case of MEM's inertial sensors, it is often indicated as a zero-g offset for accelerometers and a zero-rate offset for gyroscopes. It has been shown that there is a large range of bias variability among different commercial devices and between devices of the same model [17]. Large uncompensated bias in HAR applications can lead to difficulties in detecting states when using different devices. In these cases, oriented data fusion techniques can be used to compensate the biases' effect on the data.

Raw signal periods are further decomposed into a few numbers (in the tens) of features. These are reduced variables of original raw data that represent the main characteristics of the signal. Inertial features are usually a mixture of frequency-domain features and time-domain features, although there are some rare cases of methods that process raw accelerometer data [12]. Table 1 summarizes the most common features applied to human activity recognition using machine learning and groups them into four domain categories: statistical, frequency, time, and time–frequency. Statistical features are descriptive features that summarize and give the variability of the time series. Time-domain features give information on how inertial signals change with time. For instance, zero-crossing is the number of times the signals change from positive to negative values in a window length. Together with frequency-domain features capturing how the signal's energy is distributed over a range of frequencies, they are useful to capture the repetitive nature of a signal that often correlates to the periodic nature of the human

Table 1
Accelerometer features for machine learning applied to human activity recognition

Features	Statistical features	Kurtosis, skewness, mean, standard deviation, interquartile range, histogram, root mean square, and median absolute deviation
Time-domain features		Magnitude area, zero-crossing rate, pairwise correlation, and autocorrelation
Frequency-domain features		Energy, entropy, dominant frequency (maximum and median frequency) and power of dominant frequency, cepstral coefficients, power bandwidth, power spectral density, and fundamental frequency
Time-frequency features		Spectrogram [12], wavelets, spectral entropy [18]

activity. Their advantage is that they are usually less susceptible to signal quality variations. Time–frequency features such as spectrograms give information about the temporal evolution of the spectral content of the signals. They can represent context information in signal patterns, but they have higher computational costs than other features. Indeed, the low computational cost is a desired characteristic of HAR for their applications, and it is no surprise that most applications with smartphones that use inertial sensors use time-domain features [19].

Data Fusion

Data fusion is the concept of combining data from multiple sources to create a result with an accuracy that is higher than that obtained from a single source. IMUs can be used to simultaneously provide linear acceleration and angular velocity of the same event, as well as the device’s heading. Data fusion techniques provide complementary information to improve human activity recognition. Importantly, they can also be used to correct each other since each IMU sensor has different strengths and weaknesses that can be combined for a better solution. Accelerometers can measure gravity for long terms but are more sensitive to certain scenarios, such as spikes. Gyroscopes can be trusted for a few seconds of relative orientation changes, but the output will drift over longer time intervals, and magnetometers are less stable in environments with magnetic interferences.

Data fusion techniques can be divided into three levels of applications: sensor-level fusion, feature-level fusion, and decision-level fusion [20]. In sensor-level fusion, the raw signals from multiple sensors are combined before feature extraction. For instance, accelerometers are sensitive to sharp jerks, while gyroscopes tend to drift over the long term; thus, sensor-level fusion helps with these problems. This is achieved via signal processing algorithms, where the most popular algorithms are the Kalman filter [21] and the complementary filter [22]. The first, an iterative filter that correlates between current and previous states, consists of low- and high-pass filtering to remove gyroscope drift and accelerometer spikes. Feature fusion refers to the combination of multiple features from different sensors before entering them into a machine learning algorithm through feature selection and reduction methods such as the principal component analysis (PCA) and singular value decomposition (SVD). Feature fusion helps in identifying the correlation between features and working with a smaller set of variables. The models’ results (e.g., multiple classifiers) are combined in decision fusion to have a more accurate single decision. The aim is to implement fusion rules to get a consensus that would help in improving the algorithm’s accuracy and have a better generalization. These rules include majority voting, boosting, and stacking [23].

Benchmark Databases

In the context of HAR, there are several advantages to having access to inertial databases. The most obvious one is that it allows for comparing several solutions. Inertial databases can be used to rapidly focus on the development of signal processing and machine learning solutions before spending time on the development and deployment of hardware sensing solutions. In the past years, several public databases have appeared in the literature for smartphone and wearables studies. The list of the main databases of inertial sensors used in research work is gathered in Lima and colleague's review [19] and in Sprager and Juric's review [24].

**Global Positioning System:
Geospatial Activity**

The Global Positioning System (GPS) is a global navigation system based on a network of GPS satellites, ground control stations, and receivers that work together to determine an accurate geographic position at any point on the Earth's surface. The widespread integration of GPS into everyday objects such as smartphones, navigation systems, and wearables (GPS watches) has enabled the objective measurement of a person's location and mobility with minimal retrieval burden and recall bias [25]. At a basic level, raw data from GPS provide latitude, longitude, and time [26]. These data can be further processed to provide objective measurements of location and time, such as measurements of trajectories and locations in specific environments. Newer GPS can provide variables such as elevation, indoor/outdoor states, and speed. GPS devices have proven to be useful tools for studying and monitoring physical activity [27]. When combined with inertial sensors, it is possible to identify activity patterns and their spatial context [28]. The spatial analysis can then be contextualized with environmental attributes (presence of green space, street connectivity, cycling infrastructure, etc.). The data is often analyzed using commercial or open-source geographic information systems (GIS), software for data management, spatial analysis, and cartographic design. According to Krenn and colleagues [28], the main limitation of using GPS in health research is the loss of data quality. Indeed, urban architecture and dense vegetation can lead to signal dropouts.

2.2.2 Interaction with the Environment

When a residence uses a controller to integrate various connected objects or home automation systems, we refer to the home as a smart home. In a smart home, the role of the home controller is to integrate the home automation systems and enable them to communicate with each other. In this approach, the subject does not need to carry a device; instead, the environment is equipped with devices that can collect the required data (*see* Fig. 1d). In smart homes, we can find diverse appliances with some degree of automation. Perhaps the most popular commercial device is the smart speaker, equipped with a virtual assistant that responds to voice commands. More and more common virtual assistant technologies have expanded the use of speech processing, and the so-called vocal

biomarkers are being considered into precision medicine [29]. These technologies can be embedded in what is known as affective signal processing, for example, to monitor the mood states of home residents [30].

Smart plugs are another type of smart device that fits into existing wall outlets. They connect to the Wi-Fi or Bluetooth network and enable the control of various appliances by turning them on and off on pre-programmed schedules. Although they are not sensing devices that collect data per se, by activating the appliances, they allow the interaction between the user and the environment, and they can be used to activate sensing devices. Other smart devices that typically do not collect data from the user but enable interaction with the environment include smart light bulbs that can be turned on at specific times and allow to be controlled to create a colorful ambiance, smart thermostats to control room temperature, and smart showers. Other smart systems that allow data collection and interaction with the user and the environment include smart refrigerators that register the door's opening and the amount of food inside. They also offer the ability to view recipes and videos and adjust the water temperature through a touchscreen. Smart devices that help with sleep are smart mattresses, sleep trackers, and sleep noise machines.

Finally, when installed in strategic places, presence detectors or switches that detect the opening and closing of doors and windows can work together to create a map of presence and displacement activity inside the smart home.

The advent of smart home technology has fostered its development in medicine and human research. One such example is the use of surveillance cameras, which were initially deployed for security monitoring of goods and may now be used to detect falls by elderly persons in everyday life, thanks to advanced image processing techniques [31]. Home automation systems built around dedicated single-board computers (e.g., Raspberry Pi) expand behavioral tracking capabilities to more complex behaviors using off-the-shelf components [32].

2.2.3 Interaction Between Users

The massive adoption of Internet of Things (IoT) devices has made it possible to have a network of interconnected devices that interact to collect and analyze data using an Internet connection for remote computing (*see Fig. 1e*). Interaction between devices may be used as a proxy to inform about the collective and individual behavior of the user(s) carrying them. This interaction is possible due to numerous wireless technologies that enable communication among devices, such as Wi-Fi and Bluetooth. Moreover, a richer picture of the social world may be obtained from the traces of interactions in cyberspace, such as the analysis of individual devices' communication. Research using the phone call detail records of a sample of elderly participants in France demonstrated that such

passive data could represent a low-cost and noninvasive way to monitor the fluctuations of mood [33], working as a “social sensor containing relevant health-related insights.”

Within the wireless technologies available, Bluetooth is widely present in everyday technological devices, and it can be used as a mean to measure the interaction between users. It is based on a radio frequency that allows nearby devices to exchange data wirelessly. Bluetooth devices are paired (established logical link) before transmitting the information for security reasons. Each Bluetooth device is addressable by a unique Bluetooth device address assigned during manufacturing in addition to a textual modifiable identifier [34]. Once the devices are Bluetooth-enabled, they act as passive tools that can be used in the context of interaction monitoring between individuals. The reason why Bluetooth is better fitted to this purpose than Wi-Fi is that the former is mainly used for linking electronic devices for only short communication bouts using relatively small amounts of data and requires less power compared to Wi-Fi, which is designed to shuttle larger amounts of data between computers and the Internet. Another reason is that Bluetooth technology is rapidly evolving, offering simpler connectivity protocols between devices and better security, together with faster communication (Bluetooth V3) and lower energy consumption (Bluetooth Low Energy) with the latest version (Bluetooth 5) offering a more extensive range, speed, and bandwidth.

Implicit Bluetooth encounters can be used to passively detect implicit connections between persons, model and predict social interactions, recognize social patterns, and create networking structures without monitoring physical areas and letting people feel observed. With the COVID-19 pandemic, massive efforts to deploy contact tracing systems to notify for risk of infection used a Bluetooth protocol in smartphones as a way to identify the risk of close contact with infected individuals. In this context, Bluetooth exchanges were considered encounters [35]. This is one remarkable example of Bluetooth technology showing how it can be applied to exploit users’ interactions in real time to help manage an important health issue in modern society.

3 Applications to Brain Disorders

A growing number of applications have been developed to collect and exploit sensor data for basic science and clinical applications related to the disorders of the nervous system—as well as in human behavior in general; *see* [36]. This section presents a selection of original and representative application examples where the previously presented sensors have been put into practice to prevent, early diagnose, monitor, and create tailored help for patients, with what

is referred to as digital phenotyping. The objective of this section, far from being an extensive review of the sensor-based applications, is to give an idea to the reader of how the same sensors can be used with different objectives across a broad range of brain disorders. The brain disorders mentioned here are Alzheimer's disease (AD) (*see also* [37]), Parkinson's disease (PD) [38], epilepsy [39], multiple sclerosis [40], and some developmental disorders and psychiatric disorders [41].

3.1 Prevention

The blooming market of mobile technologies in the field of well-being and self-quantization, from basic logging to deep personal analytics, represents an opportunity to promote and assist health-enhancing behaviors. For instance, as much as 85% of US adults own a smartphone [42] and 21% an activity tracker [42]. Digital prevention uses these mobile technologies to advise and anticipate a decline in health, the goal being to prevent health threats and predict event aggravation by monitoring continuous patient status and warning indications.

An example of digital prevention in the psychiatric domain includes specific tools to prevent burnout, depression, and suicide rates. Web-based and mobile applications have been shown to be interesting tools for mitigating these severe psychiatric issues. For instance, a recent study [43] showed how the combination of a smartphone app with a wearable activity tracker was put into use to prevent the recurrence of mood disorders. With passive monitoring of the patient's circadian rhythm behaviors, their ML algorithm was able to detect irregular life patterns and alert the patients, reducing by more than 95% the amount and duration of depressive episodes, maniac or hypomanic episodes, and mood episodes.

In specific contexts known for being risk-prone with respect to mental health, e.g., high psychological demand jobs, as well as in more general professional settings, organizations have been starting to deploy workplace prevention campaigns using digital technologies [44]. In a study by Deady and colleagues [45], the authors developed a smartphone app designed to reduce and prevent depressive symptoms among a group of workers. The control group had a version of the app with a monitoring component, and the intervention group had the app version that included a behavioral activation and mindfulness intervention besides the monitoring component. Their study showed how the smartphone app helped prevent incident depression in the intervention group by showing fewer depression symptoms and less prevalence of depression over 12 months compared to the control group. Both these examples show how using smartphones and wearable devices can reduce symptoms and potentially prevent mental health decline.

3.2 Early Diagnosis

Although, in the brain care literature, most applications for diagnosis with ML use anatomical, morphological, or connectivity data derived from neuroimaging [46], there is a growing body of evidence indicating that common sensors could be used in some cases to detect behavioral and/or motor changes preceding clinical manifestations of diverse brain diseases by several years. In contrast, in neurodegenerative diseases like AD [47], PD [48], and motor neuron disease (MND) [49], the symptoms manifest when a substantial loss of neurons has already occurred, making early diagnosis challenging. Because of this, with the increasing adoption of ML in research and clinical trials, directed efforts have been made to diagnose neurodegenerative diseases early. As an example, in PD, a study used IMU in smartphones to characterize gait in the senior population, detecting gait disturbances, an early sign of PD, and showing the feasibility of the approach with a patient who showed step length and frequency disturbances and who was later formally diagnosed with PD [50]. Apathy, conventionally defined as an “absence or lack of feeling, emotion, interest or concern” [51], is one of the most frequent behavioral symptoms in neurological and psychiatric diseases. In the daily life of patients, apathy results in reduced daily activities and social interactions. These behavioral alterations may be detected as a reduction in the second-order moment (variance) of location data (as tracked with GPS measurements [52]) and in the first-order moment (average quantity) of accelerometer measures (e.g., [53] in the context of schizophrenia patients).

Sensors can also be used to differentiate between disorders that have shared symptoms, accelerating diagnosis and treatments. For instance, a study [54] that used wrist-worn devices containing accelerometers analyzed measures of sleep, circadian rhythmicity, and amplitude fluctuations to distinguish with 83% accuracy pediatric bipolar disorder (BD) and attention-deficit hyperactivity disorder (ADHD), two common psychiatric disorders that share clinical features such as hyperactivity.

3.3 Symptom and Treatment Monitoring

Monitoring day-to-day activities and the evolution of symptoms is impossible for health providers outside the clinic without automated detection of events of interest and deployment of mobile interventions. Much like apathy, described above, many other psychological constructs may be sensed from continuous monitoring of behavioral parameters, such as agitation or aberrant mobile behavior [55, 56].

Sleeping is one activity that cannot be monitored in any other way than with passive data probing in an ecological manner. Monitoring sleep is relevant when studying sleep disturbance, a core diagnostic feature of depressive disorder, anxiety disorders, bipolar disorder, and schizophrenia spectrum disorder. In this sense, sleep patterns have been scored using light sensors in mobile devices and

usage data, allowing digital phenotyping of the users that, compared to the average, go to bed and wake up later and more often. Disrupted sleep patterns have also been assessed with wrist-worn accelerometers to monitor sleep changes in various psychiatric disorders [57], as well as considered a potential psychiatric diagnostic tool in bipolar disorder, where sleep changes are a warning sign of an affective episode [58].

Given the progressive nature of some diseases, such as Alzheimer's and Parkinson's diseases, the individuals suffering from them must be monitored often or even continuously. In both cases, the patients suffer from functional and cognitive decline, where continuous objective monitoring could help detect the decline in daily capabilities providing opportunities for assistance. In the literature interested in monitoring Alzheimer's disease, the studies mainly focus on the detection of abnormal behavior, the detection of autonomy in activity performance, the provision of assistance with cognitive or memory problems, and the monitoring of functional and cognitive decline [59]. To objectively assess autonomy at home, video cameras and tags on house objects along with a mobile phone application were used in a study [60] with mild cognitively impaired patients, Alzheimer patients, and healthy controls. The activities examined included online payment, preparing a drink, medicine box preparation, and talking on the phone. To monitor cognitive decline, Lyon and colleagues from the Oregon Center for Aging and Technology (ORCATECH) [61] placed a smart sensor platform in 480 homes of an elderly population in an 8-year longitudinal study. The sensors included wireless passive infrared motion sensors, wireless magnetic contact sensors placed outside the door and in the refrigerator, a personal computer that recorded time spent in the computer and the mouse movements, worn actigraphs to measure mobility patterns, and, in some cases, connected objects such as medication trackers, phone monitors, and wireless scales. Using these multimodal data and applying sensor fusion techniques, they could identify decline in cognition, loneliness, and mood anomaly. Finally, as nighttime wanderings and memory loss are common characteristics of Alzheimer's patients, GPS solutions are increasingly used by caregivers to locate missing patients but are also recently being used in various studies [62] as effective noninvasive means of monitoring mobility in these patients. GPS solutions have also been exploited in other areas, such as in monitoring anxiety disorders. For instance, GPS data has helped predict social anxiety scores among college students by analyzing mobility features and detecting that socially anxious students avoid public areas and engage less in leisure activities to spend more time at home after school [63].

An interesting advantage of in-home monitoring of symptoms is collecting ecological data allowing clinicians to contextualize sensor data to guide potential medication changes. For instance,

Chen and colleagues [64] introduced a web-based platform that integrates data from wearable accelerometers and online surveys to estimate clinical scores of tremors, bradykinesia, and dyskinesia. The objective was to facilitate clinicians' decision-making regarding titration and timing of medications in PD patients with later-stage disease. Along the same line, in the aforementioned ORCATECH study [61], specific medication trackers (electronic pillbox) were also used to complement behavioral assessment derived from sensors: they demonstrated a significant impact of early cognitive deficits on medication adherence in everyday life.

Active probing of subjective assessment through the everyday life course of patients (commonly performed by smartphones and now smartwatches) is known as ecological momentary assessment (EMA) [61]. EMA aims at reducing memory bias and increasing the density of longitudinal data available in a single patient while exploring the possible influence of real-life contexts on cognitions and behaviors. EMA may thus capture the dynamic changes seen in psychiatric [65, 66] or neurological [67] conditions across hours, days, or longer periods, delivered according to either a predetermined schedule or in response to some event of interest, as detected by the system. EMA may also be used in combination with other passive measures and can be particularly useful to provide a ground truth concerning subjective states (e.g., mood or apathy [53]).

3.4 Tailored Help for Patients and Augmented Therapies

Personalized or precision medicine consists in using collected data to refine the diagnosis and treatment of individual patients. In this sense, connected devices and mobile technologies could contribute to tailoring patients' care. Moreover, personalized or augmented therapies can benefit from using smart devices and connected objects to add additional assistance to classic therapeutic approaches.

An example of this is epilepsy, a central system disorder that causes seizures. Not only the unpredictability of seizure occurrence is distressing for patients and contributes to social isolation, but for unattended patients with recurrent generalized tonic-clonic seizures (GTCS), this may lead to severe injuries and constitute the main risk factor of sudden unexpected death. This is why, in the epilepsy research field, much effort has been put into developing ambulatory monitoring with alarms for automated seizure detection, with most real-time application studies using wrist accelerometers, video monitoring, surface electromyography (sEMG), or under-mattress movement monitors based on electromechanical films [68]. The general purpose of using these sensors is to detect unpredictable changes in motor activity or changes in autonomic parameters characteristic of seizures.

Another illustrative case of the interest in mobile technology for helping patients in their everyday life concerns fall detection in older and /or gait-disabled persons: wireless versions of inertial and

pressure sensors have been used to monitor balance impairments in patients and to trigger an alert system when a fall is detected [69]. Data issued from mobile, wearables, and connected devices may also contribute to adjusting the therapeutic strategy followed by the healthcare provider. Omberg and colleagues [70] demonstrated that in Parkinson's disease patients, remote assessment through smartphones correlated with in-clinic evaluation of disease severity. In the context of rehabilitation following cerebrovascular lesions or neurocognitive training in neuropsychiatric disorders, connected devices may also contribute to making the rehabilitation/training program more engaging for patients and improving its real-life efficacy [71].

Finally, in the context of psychiatric disorders, mobile technologies may also support ecological momentary or just-in-time interventions (EMI), a promising venue for augmenting mental healthcare and psychotherapy through digital technologies [72, 73].

4 Considerations and Challenges

When conceptualizing and developing a project involving human behavior recognition, it is essential to anticipate the known challenges and difficulties that can be encountered. We present the general known common challenges for connected devices under three groups: (1) those that are related to sensor function per se, (2) the challenges related to the signal processing and machine learning methods used to exploit the data and that are partly shared with other pattern recognition fields, and (3) the challenges raised by deploying real-life applications.

4.1 Related to Sensor Function

4.1.1 Sample Rate Stability

We refer to sample rate stability as the homogenous regularity time spans between consecutive samples. In a reliable device, the difference between different time spans between successive measurements is close to zero. When this is not the case, the true measure by the sensor and the timestamp registered by the application differs. Common sources of sample rate instability are the inherent jitter by non-real-time operating systems that cannot guarantee critical execution time or access to resources and the additional communication delay between the devices and applications.

4.1.2 The Choice of Technology

Sensors are usually input devices that take part in a bigger system, sending information to a processing unit so that the signals can be analyzed. When choosing a technology to work with, a careful choice of all of the parts must be pre-studied to avoid issues in usability and signal quality since these will have an impact on the difficulty of development and deployment, as well as on the long-term use of the technology. For instance, if we need to record

inertial measurements and the body location is not a major issue, deciding between a dedicated IMU device, a smartwatch, or a smartphone would be necessary. Smartwatches, having fewer resources than smartphones, show larger sampling instabilities, especially under high CPU load [17], and then the question would be if a smartwatch would then be appropriate for the application, and so, what model would provide the best sampling stability over long recordings? Hardware memory usage limitations and power consumption are critical criteria to consider, especially for the long-term use of connected devices. Another issue is the open access to commercial devices. Most commercial devices (smartphones, smartwatches, and connected devices) offer the developers the opportunity to use their integrated sensors to develop applications using their platforms (i.e., Android, iOS, Tizen, etc.). Usually, the development of these commercial devices comes with certain restrictions. For instance, the developers do not have complete access to the device, and to modifications of the operating system, the programming language is usually restricted, and some pre-programmed tasks are usually impossible to modify.

4.1.3 Power Consumption

One of the main problems preventing the massive expansion and adoption of HAR applications is excessive battery power consumption [75]. Indeed, the major problems that lead to data loss are empty batteries, where the main sources of high power consumption are the high data processing load and the continuous use of sensors. Some strategies can be adopted to minimize energy consumption, although these imply a tradeoff between energy consumption, signal richness, and the accuracy of classification models. The first strategy consists of on-demand activation of sensors only when necessary, in contrast to continuous sampling; this requires a continuous supplementary routine that automatically determines when the timing is appropriate to interrogate the sensor(s). Tech companies have dealt with this problem by integrating “sensor hubs,” i.e., low-power coprocessors that are dedicated to reading, buffering, and processing continuous sensor data for specific functions such as step counting and spoken word detection (for instance, the specific function of detecting the famous popular voice commands “hello google” or “Alexa” for Google’s and Amazon’s vocal assistants). The second strategy consists of choosing the sampling frequency of data collection. The higher the frequency in sampling data, the more energy the sensors, the processor, and the memory unit use. Previous knowledge of the signals and the frequency necessary to capture events is needed to select a sampling frequency which is a good tradeoff between capturing relevant signal information and avoiding an unnecessary battery drop. The third strategy focuses on the applications where the data is processed on the device by strategically selecting lightweight features

to reduce the data processing load. For instance, in inertial data processing, time-domain features have lower computational costs than frequency- and time-frequency-domain features. Considering how sensors' power consumption and applications affect battery life in worn systems with small batteries is essential. Since total power load is hard to estimate, it depends on many external factors such as the main application processor, access to memory by other applications, etc. A good practice is to record battery statistics for several days across different participants to estimate real-life use and average battery life.

4.2 Related to the Data

4.2.1 Data Collection

Data collection consists of data acquisition, data labeling, and existing database improvement. It is one critical challenge in machine learning and often the most time-consuming step in an end-to-end machine learning application due to the time spent collecting the data, cleaning, labeling (for supervised learning), and visualizing it. The data required by the machine learning models can be experimental, retrospective, observational, and, in some cases, synthetic data. While retrospective data collection methods such as surveys and interviews are easy to deploy, they are subject to recall and to self-selection bias, and they might add tedious collection logistical issues if tools and programs in mobile devices are not deployed. Retrospective data collection is sometimes the only means to capture subjective experiences in daily life. Observation methods such as video-camera surveillance can be impractical for large-scale deployment and are often primarily used in small sample applications. Generation of synthetic data is sometimes necessary to overcome the lack of data in some domains, notably annotated medical data. This kind of data is created to improve AI models through data augmentation from models that simulate outcomes given specific inputs such as bio-inspired data [74], physical simulations, or AI-driven generative models [75]. The issue with this is that there is a lack of regulatory frameworks involving synthetic data and their monitoring. Their evaluation could be done with a Turing test, yet this may be prone to inter- and intra-observer variabilities. Plus, data curation protocols can be as tedious and laborious as collecting and labeling real data.

The availability of large-scale, curated scientific datasets is crucial for developing helpful machine learning benchmarks for scientific problems [76], especially for supervised learning solutions where data volume and modality are relevant [77]. Even though machine learning has been used in many domains, there is still a broad panel of applications and fields, such as neuroscience and psychiatry, with few or even nonexistent training databases. This is the case for connected devices' and sensors' derived datasets for brain disorder research. In contrast, there are nowadays larger neuroimaging and biological databases available, e.g., the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Allen Brain

Atlas. Fortunately, following the moderate adoption of machine learning in the brain research field, a trend toward increasing sharing of resources has emerged, but for now, it is mainly in the neuroimaging field. Each year, more scientific open data becomes available, although their curation, maintenance, and distribution for public consumption are challenging, especially for large-scale datasets. Another increasing trend in data collection or human annotation of data is through crowdsourcing marketplaces, having the advantage of giving access to diverse profiles from a large population sample, enabling to find more representative examples to train the models.

4.2.2 Database Validation and Signal Richness

For applications in medicine and healthcare, the datasets used to train the ML models should undergo detailed examination because they are central to understanding the model's biases and pitfalls. Before adopting an openly available dataset or creating one, there are some considerations that we have to keep in mind. Firstly, ensure a minimal chance of sample selection biases in the database (for instance, data acquired with particular equipment or with a particular setting). Errors from sample selection biases become evident when the model is deployed in settings different from those used for training. Secondly, we must be aware of the class imbalance problem that often occurs in cases where the data is rare (for instance, in low samples associated with rare diseases), which could negatively affect models designed for prognosis and early diagnosis. A few techniques can be adopted to help with class imbalance, such as resampling, adding synthetic data, or working directly with the model, such as weighting the cost function of neural networks.

The data, often obtained from scientific experiments, should be rich enough to allow different analysis and exploration methods and carefully labeled when required. For instance, a semantic discrepancy in the labels can dilute the training pool and confuse the classifier [78, 79]. In contrast with free-form text or audio to mark the activities, imperfect labeling by the users can occur when scoring the samples with fixed labels. For instance, labeling similar activities from IMU systems, such as running and jogging under the fixed label-running, can induce errors in feature extraction because of their interactivity similarity.

It is also important for signal richness to consider subject variability and consider differences between gender, age, and any other characteristic that could lead to improper data representation. Naïve assumptions can cause actual harm by stigmatizing a population subgroup when there is an implicit bias in data collection, selection, and processing [80]. These can be addressed by expanding the solutions to inclusion at all levels and carefully auditing all stages of the development pipeline.

4.3 Deployment for Real-Life Applications

Real-work applicability requires that the accuracy be tested off laboratory settings, considering real-life factors besides technology function and data collection. Before deploying a solution using connected objects and sensors, these real-life use considerations should be addressed without depreciation. Indeed, factors such as user acceptance and behavior around the devices might even be more important than having a high positive prospect of technology. Here we briefly present three factors to keep in mind. These include thinking ahead of privacy issues and how to handle them, the potential degree of adoption, and wearability and instrumentation unobtrusiveness.

4.3.1 Health Privacy

Using mobile devices, connected objects, and sensors to collect data for machine learning for health applications is a process that generates data from human lives. In this sense, privacy is a common concern with health data. The concept of privacy in health refers to the contextual rules around generated data or information: how it flows depending on the actors involved, what is the process by which it is accessed, the frequency of the access, and the purpose of the access [81].

The machine learning community has generally valued and embraced the concept of openness. It is common for code and datasets to be publicly released and paper preprints to be available on dedicated archival services before an article is published (despite rejection). Therefore, regulatory bodies should encourage and enforce data holders to collect and provide data under clear legal protection. To ensure data security, these regulations might suggest adopting different solutions: not transmitting raw data, having an isolated sensor network, transmitting encrypted data, and controlling data access authorizations [82]. While individual countries decide where to draw the line regarding regulations, sometimes, depending on the data type, this is more or less difficult to define. For instance, there might be clearer limits on the exploitation and use of patient video recordings because there is explicit reasoning that the patient's identity is easily accessible with image processing. In contrast, this reasoning is less straightforward with other types of data. For instance, even though inertial sensor data might be sufficient to obtain information about a person based on their biometric movement patterns, these sensors are currently not perceived as particularly sensitive by the public. Part of this is because their privacy implications are less well-understood [83]. Thus, they tend to be much less protected (e.g., in wearable devices and mobile apps) compared to other sensors such as GPS, cameras, and microphones. Therefore, requiring proper permission, conscious advertised participation, and explicit consent from the user is essential, no matter the nature of the data collected.

4.3.2 Perception and Adoption

The perception and adoption of mobile devices, connected objects, and sensors refer to the negative or positive way the deployed solutions are regarded, understood, and interpreted by the users. This degree of perception directly affects the adherence to a protocol and the solution's use or adoption over the long term. It is one of the most important factors to evaluate ahead of deployment in real-life scenarios. It implies a conscious effort to understand the patient's situation and point of view. It can be overseen by developers and researchers who could focus more on the technical or scientific challenges to overcome or who, because of naivety or distance to the patient's reality, might unwarily not include these considerations in their designs.

The Technology Acceptance Model (TAM) [84], which can be applied to mobile devices, connected objects, and sensors, postulates that two factors predict technology acceptance. The first one is the perceived usefulness or the degree to which a person believes a particular solution will enhance or improve the performance of a specific task. The second factor is perceived ease of use or the degree to which a person believes the solution proposed will be free of effort. The perceived ease of use and usefulness might vary according to the population target and should be studied carefully before deployment. For instance, the perceived ease of use is essential for the elderly [85], who are not core consumers of mobile wireless healthcare technology. There are, of course, other models and theories [86, 87] that have been published since the TAM was proposed, and they include other essential factors to take into account, such as social influence, performance and effort expectancy, and facilitating conditions, or the perceptions of the resources and support available. Although these models have several limitations [88], the identified factors are a good starting point to consider when designing a solution involving wearable, mobile devices, and connected objects. In addition to those factors, clear limits in the cost and benefit ratio of the technology must be communicated since it is one of the main barriers to their acceptance. In that sense, the scientific and healthcare community is responsible for efficiently approaching the patients and clearly explaining the expected positive outcomes and the advantages and disadvantages of the device's ecosystems.

4.3.3 Wearability and Instrumentation Unobtrusiveness

Wearability refers to the locations where the sensors are placed and how they are attached to those locations. Wearable devices are typically attached to the body or embedded in clothes and accessories. They are smartwatches and bracelets for activity trackers, smart jewelry, smart clothing, head-mounted devices, and ear devices [89]. Wearability is an aspect to consider because of its direct impact on data collection, signal richness, and quality. The goal is to ensure the device's prolonged and correct use.

In the 1990s, Gemperle and colleagues [90] proposed the first ergonomic guidelines on wearability. Since then, different “wearability maps” have been proposed to approximate the best unobtrusive locations for sensor placement in the human body. A source of the problem in wearability is that the sensors should be securely attached to the human body to prevent relative motion, signal artifacts, and degraded sensing accuracy. Smartwatches are desired to be worn on the dominant arm to capture most of the hand movement, but it is more comfortable for people to wear them on the passive arm.

Similar to wearable devices, one desired characteristic of deployed sensors and tags in the environment is to ensure unobtrusiveness. Unobtrusive sensing allows continuous recording of the patient’s activities, behaviors, and physiological parameters without inconveniences to everyday life [82]. This can be achieved by embedding small objects interacting with the subject into the ambient environment, for which the design and usability [91], especially for long-term monitoring, have been considered. There are some devices that are perceived as more invasive than others. For instance, special measures are taken when using cameras regarding sensor selection and sensor placement [82].

4.4 Incorporation into Clinical Care

Although there is great potential for connected devices and sensors to prevent, early diagnose, monitor, and create tailored help for patients suffering from brain diseases, there is still a gap to fill to drive transformational changes in health. Besides the challenges mentioned in this section, significant barriers to clinical adoption include the lack of evidence in support of clinical use, the rapid technological development and obsolescence, and the lack of reimbursement models. These problems are often highlighted in preliminary reports of government proposals [92–94] and publications related to mobile health challenges [95, 96].

There is a need for an extensive collection of real-world patient-generated data to reinforce clinical evidence that will change health-care delivery. To date, there is a limitation due to an underpowered number of available pilot datasets that make the comparability of studies difficult and therefore the adoption of these new technologies into the clinical field. Indeed, sensor datasets come mainly from actigraphy and are not as numerous as available neuroimaging, MEG, or EEG datasets.

Opposite to the few large patient-generated evidence, the number of solutions for connected devices and sensors with added features continues to grow every year. This rapid development of technologies represents a challenge to clinicians who might perceive difficulty in the feasibility and scalability of real-word implementations within the clinical workflow, especially since it is noticeable that devices become obsolete, outdated, or no longer useful very quickly. Another negative impact of the higher number

of alternatives in the market is that too much choice can be overwhelming. In clinical trials and research, it can be challenging to choose a technical solution when there is little or no clinical evidence and when the features proposed differ significantly between solutions. Even with well-established companies, for the consumers, there is no guarantee that a product or its support will not be discontinued in the short term or that the product will not be rapidly replaced with a newer model. At the same time, ensuring that the chosen product will be well integrated with other products (e.g., compatible bricks between other sensors, software, operating systems, and processing units) is challenging. These factors add up to the paradox of choice [97], and it is a known consequence of choice overload.

With newer connected objects and sensors that appear in the market every month, there is also a rise in their associated mobile applications available. Among these mobile applications, the most popular categories are sports and fitness activity trackers, diet and nutrition, weight loss coaching, stress reduction and relaxation, menstrual period and pregnancy tracking, hospital or medical appointment tracking, patient community, and telemedicine [98]. Most of these applications are not regulated medical health solutions that work with certified medical devices. They are dedicated to consumers only (not intended for collaboration between patients and healthcare professionals) and are usually considered or displayed as well-being apps. In this sense, while various governments worldwide have opted for different lines of action regarding the consideration of connected objects and sensors in their health programs, the appropriate reimbursement models in place are far from being well integrated into regulatory norms. Take the example of France, where connected objects are rarely reimbursed by social security. For a product to be prescribed by a physician, it must be considered a regulatorily approved medical device, i.e., be registered in an official list of medical services and products. This list also establishes the proper use of the device, the support cost, the characteristics of the product, and the number of possible prescription renewals. The heavy administrative burden required to get registered discourages potential players from requesting medical approval. In particular, the product has to meet several compliance rules of the High Authority of Health (HAS), including the proven good performance of the connected object, the reliability of the medical data transmitted, and the respect and protection of personal and confidential data.

Even though many available connected objects and their mobile applications are not regulated medical health solutions, their rapid spread and adoption among the public are starting to pave the way for motivating future democratization and integration of these devices in public health policies.

5 Discussion

With the amount of innovation and development of smart devices and connected objects, together with the widespread of ML algorithms implemented in faster processing units, we are now many steps closer to having a better understanding of the underlying neural mechanisms of brain disorders with the hope to better intervene at different stages: by preventing health decline, by early and more accurately diagnosing, and by helping to better treat and monitor patients.

In this chapter, we presented the different types of data that one can gather with these devices according to the passive or active role that the user takes in their collection. Many of them are now widely adopted by modern society and used for self-monitoring (e.g., fitness trackers containing IMUs) or in smart home settings (e.g., virtual assistants and presence detectors). When these devices are used together, they represent an opportunity for data fusion allowing the joint analysis of multiple datasets that provide an enhanced complementary view of the phenomenon of interest (e.g., detecting a compulsive behavior like handwashing by combining inertial and acoustic data from a smartwatch). Without a doubt, some brain disorders are better suited for sensor-based assessments, like PD, because of their prominent motor symptoms, unlike other brain disorders whose symptom assessment requires the combination of close behavior observation and access to mental insight (e.g., mood disorders). In the second case, combining sensor data would reduce uncertainty in monitoring and diagnosing, especially when the samples are taken continuously in an ecological manner.

Despite the promising results obtained with these intelligent systems, several conditions need to be addressed before a lab-made application becomes integrated into the clinical routine and in an unsupervised domestic environment. Indeed, most publications do not reach the final phase to be considered as medical devices. Concerning the use of sensors and devices for data collection, a series of considerations to be regarded was presented in Subheading 4. Even though this list could be extended, overall, the main goal remains to assure reproducibility and unbiased collection of high-quality data since ML models can only go so far as the data they rely on.

An exciting, promising extension of the capabilities of smart devices and connected objects is their integration in a closed-loop setting, where the devices serve as real-time continuous monitoring tools that respond to events of interest to treat or intervene on demand and in real time. Indeed, this is a promising approach because of the advantage of early intervention.

Furthermore, we are currently experiencing a new medical revolution with new sensors. Besides what has been presented here, much effort has been put into developing wearable biosensors. These are sensing devices that recognize biological elements (e.g., enzymes, antibodies, and cell receptors), the most known example being glucose monitoring devices. These bioreceptor units are still in their infancy in terms of use and acceptance by the neuroscientific field and medical community in general, but we anticipate that their use and development will continue to grow in the brain disorder research field as smart devices and connected objects have.

Finally, as data and better processing techniques keep increasing, more collaborations between engineers, researchers, and clinicians are formed to contribute to the field of brain disorders positively. We believe that, in the foreseeable future, the rapid evolution of the presented technologies, their use, and their adoption will be key to revolutionizing and addressing the challenges of the traditional medical approach regarding brain disorders.

Acknowledgments

This work was supported by the following grants (SLMG, EB): Agence Nationale de la Recherche (ANR-19-ICM-DOPALOOPS, ANR-22-ICM-PREDICTOC) and Fondation de France.

The authors would like to thank Dr. Renaud David for reviewing this chapter and providing insightful comments.

References

- Sim I (2019) Mobile devices and health. *N Engl J Med* 381(10):956–968. <https://doi.org/10.1056/NEJMra1806949>
- Laput G et al (2021) Methods and apparatus for detecting individual health related events. US20210063434A1, 04 Mar 2021. Accessed: 04 Oct 2022. [Online]. Available: <https://patents.google.com/patent/US20210063434A1/en>
- Merkel S, Kucharski A (2019) Participatory design in gerontechnology: a systematic literature review. *The Gerontologist* 59(1):e16–e25. <https://doi.org/10.1093/geront/gny034>
- SENSE-PARK Consortium et al (2015) Participatory design in Parkinson’s research with focus on the symptomatic domains to be measured. *J Parkinsons Dis* 5(1):187–196. <https://doi.org/10.3233/JPD-140472>
- Thabrew H, Fleming T, Hetrick S, Merry S (2018) Co-design of eHealth interventions with children and young people. *Front Psychiatry* 9. Accessed: 04 Oct 2022. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00481>
- Szul MJ, Bompas A, Sumner P, Zhang J (2020) The validity and consistency of continuous joystick response in perceptual decision-making. *Behav Res Methods* 52(2):681–693. <https://doi.org/10.3758/s13428-019-01269-3>
- Li X, Liang Z, Kleiner M, Lu Z-L (2010) RTbox: a device for highly accurate response time measurements. *Behav Res Methods* 42(1):212–225. <https://doi.org/10.3758/BRM.42.1.212>
- Spivey MJ, Dale R (2006) Continuous dynamics in real-time cognition. *Curr Dir Psychol Sci* 15(5):207–211. <https://doi.org/10.1111/j.1467-8721.2006.00437.x>
- Piwek L, Ellis DA, Andrews S, Joinson A (2016) The Rise of Consumer Health Wearables: Promises and Barriers. *PLoS Med* 13(2):

- e1001953. <https://doi.org/10.1371/journal.pmed.1001953>
10. Faisal S, Ivo J, Patel T (2021) A review of features and characteristics of smart medication adherence products. *Can Pharm J (Ott)* 154(5):312–323. <https://doi.org/10.1177/17151635211034198>
 11. Laine C, Davidoff F (1996) Patient-centered medicine: a professional evolution. *JAMA* 275(2):152–156. <https://doi.org/10.1001/jama.1996.03530260066035>
 12. Laput G, Harrison C (2019) Sensing fine-grained hand activity with smartwatches. In: Proceedings of the 2019 CHI conference on human factors in computing systems, Glasgow Scotland UK, May 2019, pp 1–13. <https://doi.org/10.1145/3290605.3300568>
 13. Yang C-C, Hsu Y-L (2010) A review of accelerometry-based wearable motion detectors for physical activity monitoring. *Sensors* 10(8):7772–7788. <https://doi.org/10.3390/s100807772>
 14. Jones PJ et al (2021) Feature selection for unsupervised machine learning of accelerometer data physical activity clusters – a systematic review. *Gait Posture* 90:120–128. <https://doi.org/10.1016/j.gaitpost.2021.08.007>
 15. Bouten CVC, Koekkoek KTM, Verduin M, Kodde R, Janssen JD (1997) A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. *IEEE Trans Biomed Eng* 44(3):136–147. <https://doi.org/10.1109/10.554760>
 16. Laput G, Xiao R, Harrison C (2016) ViBand: high-fidelity bio-acoustic sensing using commodity smartwatch accelerometers. In: Proceedings of the 29th annual symposium on user interface software and technology, New York, NY, USA, Oct 2016, pp 321–333. <https://doi.org/10.1145/2984511.2984582>
 17. Stisen A et al (2015) Smart devices are different: assessing and mitigating mobile sensing heterogeneities for activity recognition. In: Proceedings of the 13th ACM conference on embedded networked sensor systems, New York, NY, USA, Nov 2015, pp 127–140. <https://doi.org/10.1145/2809695.2809718>
 18. Khan AM, Lee YK, Lee SY (2010) Accelerometer's position free human activity recognition using a hierarchical recognition model. In: The 12th IEEE international conference on e-health networking, applications and services, July 2010, pp 296–301. <https://doi.org/10.1109/HEALTH.2010.5556553>
 19. Sousa Lima W, Souto E, El-Khatib K, Jalali R, Gama J (2019) Human activity recognition using inertial sensors in a smartphone: an overview. *Sensors* 19(14):14. <https://doi.org/10.3390/s19143213>
 20. Webber M, Rojas RF (2021) Human activity recognition with accelerometer and gyroscope: a data fusion approach. *IEEE Sensors J* 21(15):16979–16989. <https://doi.org/10.1109/JSEN.2021.3079883>
 21. Castanedo F (2013) A review of data fusion techniques. *Sci World J* 2013:e704504. <https://doi.org/10.1155/2013/704504>
 22. Islam T, Islam MS, Shahid-Ul-Mahmud M, Hossam-E-Haider M (2017) Comparison of complementary and Kalman filter based data fusion for attitude heading reference system. *AIP Conf Proc* 1919(1):020002. <https://doi.org/10.1063/1.5018520>
 23. Nweke HF, Teh YW, Mujtaba G, Al-garadi MA (2019) Data fusion and multiple classifier systems for human activity detection and health monitoring: review and open research directions. *Inf Fusion* 46:147–170. <https://doi.org/10.1016/j.inffus.2018.06.002>
 24. Sprager S, Juric M (2015) Inertial sensor-based gait recognition: a review. *Sensors* 15(9):22089–22127. <https://doi.org/10.3390/s150922089>
 25. Breasail MÓ et al (2021) Wearable GPS and accelerometer technologies for monitoring mobility and physical activity in neurodegenerative disorders: a systematic review. *Sensors* 21(24):24. <https://doi.org/10.3390/s21248261>
 26. Jankowska MM, Schipperijn J, Kerr J (2015) A framework for using GPS data in physical activity and sedentary behavior studies. *Exerc Sport Sci Rev* 43(1):48–56. <https://doi.org/10.1249/JES.000000000000035>
 27. Maddison R, Ni Mhurchu C (2009) Global positioning system: a new opportunity in physical activity measurement. *Int J Behav Nutr Phys Act* 6(1):73. <https://doi.org/10.1186/1479-5868-6-73>
 28. Krenn PJ, Titze S, Oja P, Jones A, Ogilvie D (2011) Use of global positioning systems to study physical activity and the environment: a systematic review. *Am J Prev Med* 41(5):508–515. <https://doi.org/10.1016/j.amepre.2011.06.046>
 29. Fagherazzi G, Fischer A, Ismael M, Despotovic V (2021) Voice for health: the use of vocal biomarkers from research to clinical practice. *Digit Biomark* 5(1):78–88. <https://doi.org/10.1159/000515346>
 30. Fedotov D, Matsuda Y, Minker W (2019) From smart to personal environment: integrating emotion recognition into smart houses. In:

- 2019 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops), Mar 2019, pp 943–948. <https://doi.org/10.1109/PERCOM.2019.8730876>
31. De Miguel K, Brunete A, Hernando M, Gambao E (2017) Home camera-based fall detection system for the elderly. *Sensors* 17(12):12. <https://doi.org/10.3390/s17122864>
 32. Koutli M, Theologou N, Tryferidis A, Tzovaras D (2019) Abnormal behavior detection for elderly people living alone leveraging IoT sensors. In: 2019 IEEE 19th international conference on Bioinformatics and Bioengineering (BIBE), Oct 2019, pp 922–926. <https://doi.org/10.1109/BIBE.2019.900173>
 33. Aubourg T, Demongeot J, Renard F, Provost H, Vuillerme N (2019) Association between social asymmetry and depression in older adults: a phone Call Detail Records analysis. *Sci Rep* 9(1):1. <https://doi.org/10.1038/s41598-019-49723-8>
 34. Davies N, Friday A, Newman P, Rutledge S, Storz O (2009) Using bluetooth device names to support interaction in smart environments. In: Proceedings of the 7th international conference on mobile systems, applications, and services, New York, NY, USA, pp 151–164. <https://doi.org/10.1145/1555816.1555832>
 35. Barthe G et al (2022) Listening to bluetooth beacons for epidemic risk mitigation. *Sci Rep* 12(1):1. <https://doi.org/10.1038/s41598-022-09440-1>
 36. Box-Steffensmeier JM et al (2022) The future of human behaviour research. *Nat Hum Behav* 6(1):15–24. <https://doi.org/10.1038/s41562-021-01275-6>
 37. Kourtis LC, Regele OB, Wright JM, Jones GB (2019) Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *NPJ Digit Med* 2(1):9. <https://doi.org/10.1038/s41746-019-0084-2>
 38. Channa A, Popescu N, Ciobanu V (2020) Wearable solutions for patients with Parkinson's disease and neurocognitive disorder: a systematic review. *Sensors* 20(9):2713. <https://doi.org/10.3390/s20092713>
 39. Asadi-Pooya AA, Mirzaei Damabi N, Rostaminejad M, Shahsavandi M, Asadi-Pooya A (2021) Smart devices/mobile phone in patients with epilepsy? A systematic review. *Acta Neurol Scand* 144(4):355–365. <https://doi.org/10.1111/ane.13492>
 40. Marziniak M, Brichetto G, Feys P, Meyding-Lamadé U, Vernon K, Meuth SG (2018) The use of digital and remote communication technologies as a tool for multiple sclerosis management: narrative review. *JMIR Rehabil Assist Technol* 5(1):e7805. <https://doi.org/10.2196/rehab.7805>
 41. Torous J, Onnella J-P, Keshavan M (2017) New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. *Transl Psychiatry* 7(3):e1053. <https://doi.org/10.1038/tp.2017.25>
 42. Pew Research Center (2021) Mobile fact sheet. Pew Research Center: Internet, Science & Tech. 07 Apr 2021. <https://www.pewresearch.org/internet/fact-sheet/mobile/> (accessed 06 Oct 2022).
 43. Cho C-H et al (2020) Effectiveness of a smartphone app with a wearable activity tracker in preventing the recurrence of mood disorders: prospective case-control study. *JMIR Ment Health* 7(8):e21283. <https://doi.org/10.2196/21283>
 44. Brassey J, Guntner A, Isaak K, Silberzahn T (2021) Using digital tech to support employees' mental health and resilience. McKinsey & Company
 45. Deady M et al (2022) Preventing depression using a smartphone app: a randomized controlled trial. *Psychol Med* 52(3):457–466. <https://doi.org/10.1017/S0033291720002081>
 46. Vogels EA (2020) About one-in-five Americans use a smart watch or fitness tracker. Pew Research Center. 09 Jan 2020. <https://www.pewresearch.org/fact-tank/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/> (accessed 06 Oct 2022).
 47. Donev R, Kolev M, Millet B, Thome J (2009) Neuronal death in Alzheimer's disease and therapeutic opportunities. *J Cell Mol Med* 13(11–12):4329–4348. <https://doi.org/10.1111/j.1582-4934.2009.00889.x>
 48. Michel PP, Hirsch EC, Hunot S (2016) Understanding dopaminergic cell death pathways in Parkinson disease. *Neuron* 90(4):675–691. <https://doi.org/10.1016/j.neuron.2016.03.038>
 49. Boillée S, Vande Velde C, Cleveland DW (2006) ALS: a disease of motor neurons and their nonneuronal neighbors. *Neuron* 52(1):39–59. <https://doi.org/10.1016/j.neuron.2006.09.018>
 50. Lan K-C, Shih W-Y (2014) Early diagnosis of Parkinson's disease using a smartphone. *Procedia Comput Sci* 34:305–312. <https://doi.org/10.1016/j.procs.2014.07.028>

51. Levy R, Dubois B (2006) Apathy and the functional anatomy of the prefrontal cortex–basal ganglia circuits. *Cereb Cortex* 16(7):916–928. <https://doi.org/10.1093/cercor/bhj043>
52. Saeb S et al (2015) Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J Med Internet Res* 17(7):e175. <https://doi.org/10.2196/jmir.4273>
53. Kluge A et al (2018) Combining actigraphy, ecological momentary assessment and neuroimaging to study apathy in patients with schizophrenia. *Schizophr Res* 195:176–182. <https://doi.org/10.1016/j.schres.2017.09.034>
54. Faedda GL et al (2016) Actigraph measures discriminate pediatric bipolar disorder from attention-deficit/hyperactivity disorder and typically developing controls. *J Child Psychol Psychiatry* 57(6):706–716. <https://doi.org/10.1111/jcpp.12520>
55. Favela J, Cruz-Sandoval D, Morales-Tellez A, Lopez-Navia IH (2020) Monitoring behavioral symptoms of dementia using activity trackers. *J Biomed Inform* 109:103520. <https://doi.org/10.1016/j.jbi.2020.103520>
56. Manley NA et al (2020) Long-term digital device-enabled monitoring of functional status: implications for management of persons with Alzheimer's disease. *Alzheimers Dement Transl Res Clin Interv* 6(1):e12017. <https://doi.org/10.1002/trc2.12017>
57. Wainberg M et al (2021) Association of accelerometer-derived sleep measures with lifetime psychiatric diagnoses: a cross-sectional study of 89,205 participants from the UK Biobank. *PLoS Med* 18(10):e1003782. <https://doi.org/10.1371/journal.pmed.1003782>
58. Fellendorf FT et al (2021) Monitoring sleep changes via a smartphone app in bipolar disorder: practical issues and validation of a potential diagnostic tool. *Front Psychiatry* 12:641241. <https://doi.org/10.3389/fpsyg.2021.641241>
59. Gillani N, Arslan T (2021) Intelligent sensing technologies for the diagnosis, monitoring and therapy of Alzheimer's disease: a systematic review. *Sensors* 21(12):4249. <https://doi.org/10.3390/s21124249>
60. Karakostas A et al (2020) A French-Greek cross-site comparison study of the use of automatic video analyses for the assessment of autonomy in dementia patients. *Biosensors* 10(9):E103. <https://doi.org/10.3390/bios10090103>
61. Lyons BE et al (2015) Pervasive computing technologies to continuously assess Alzheimer's disease progression and intervention efficacy. *Front Aging Neurosci* 7:102. <https://doi.org/10.3389/fnagi.2015.00102>
62. Cullen A, Mazhar MKA, Smith MD, Lithander FE, Breasail MO, Henderson EJ (2022) Wearable and portable GPS solutions for monitoring mobility in dementia: a systematic review. *Sensors* 22(9):3336. <https://doi.org/10.3390/s22093336>
63. Boukhechba M, Chow P, Fua K, Teachman BA, Barnes LE (2018) Predicting social anxiety from global positioning system traces of college students: feasibility study. *JMIR Ment Health* 5(3):e10101. <https://doi.org/10.2196/10101>
64. Chen B-R et al (2011) A web-based system for home monitoring of patients with Parkinson's disease using wearable sensors. *IEEE Trans Biomed Eng* 58(3):831–836. <https://doi.org/10.1109/TBME.2010.2090044>
65. Morgièvre M et al (2020) A digital companion, the Emma app, for ecological momentary assessment and prevention of suicide: quantitative case series study. *JMIR MHealth UHealth* 8(10):e15741. <https://doi.org/10.2196/15741>
66. Seppälä J et al (2019) Mobile phone and wearable sensor-based mHealth approaches for psychiatric disorders and symptoms: systematic review. *JMIR Ment Health* 6(2):e9819. <https://doi.org/10.2196/mental.9819>
67. Cain AE, Depp CA, Jeste DV (2009) Ecological momentary assessment in aging research: a critical review. *J Psychiatr Res* 43(11):987–996. <https://doi.org/10.1016/j.jpsychires.2009.01.014>
68. Rugg-Gunn F (2020) The role of devices in managing risk. *Epilepsy Behav* 103. <https://doi.org/10.1016/j.yebeh.2019.106456>
69. Usmani S, Saboor A, Haris M, Khan MA, Park H (2021) Latest research trends in fall detection and prevention using machine learning: a systematic review. *Sensors* 21(15):5134. <https://doi.org/10.3390/s21155134>
70. Omberg L et al (2022) Remote smartphone monitoring of Parkinson's disease and individual response to therapy. *Nat Biotechnol* 40(4):4. <https://doi.org/10.1038/s41587-021-00974-9>
71. Robert P et al (2021) Efficacy of serious exergames in improving neuropsychiatric symptoms in neurocognitive disorders: results of the X-TORP cluster randomized trial. *Alzheimers Dement Transl Res Clin Interv* 7(1). <https://doi.org/10.1002/trc2.12149>
72. Balaskas A, Schueler SM, Cox AL, Doherty G (2021) Ecological momentary interventions for mental health: a scoping review. *PLoS One*

- 16(3):e0248152. <https://doi.org/10.1371/journal.pone.0248152>
73. Stern E et al (2022) How can digital mental health enhance psychiatry? *Neuroscientist* 10738584221098604. <https://doi.org/10.1177/10738584221098603>
74. Mondragón-González SL, Burguière E (2017) Bio-inspired benchmark generator for extracellular multi-unit recordings. *Sci Rep* 7(1):1. <https://doi.org/10.1038/srep43253>
75. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F (2021) Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng* 5(6):6. <https://doi.org/10.1038/s41551-021-00751-8>
76. Thiyyagalingam J, Shankar M, Fox G, Hey T (2022) Scientific machine learning benchmarks. *Nat Rev Phys* 4(6):6. <https://doi.org/10.1038/s42254-022-00441-7>
77. Myszczynska MA et al (2020) Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol* 16(8):440–456. <https://doi.org/10.1038/s41582-020-0377-8>
78. Abdullah S, Lane N, Choudhury T (2012) Towards population scale activity recognition: a framework for handling data diversity. *Proc AAAI Conf Artif Intell* 26(1):851–857
79. Peebles D, Lu H, Lane N, Choudhury T, Campbell A (2010) Community-guided learning: Exploiting mobile sensor users to model human behavior. *Proc AAAI Conf Artif Intell* 24(1):1600–1606. <https://doi.org/10.1609/aaai.v24i1.7731>
80. Ghassemi M, Mohamed S (2022) Machine learning and health need better values. *NPJ Digit Med* 5(1):1. <https://doi.org/10.1038/s41746-022-00595-9>
81. Price WN, Cohen IG (2019) Privacy in the age of medical big data. *Nat Med* 25(1):37–43. <https://doi.org/10.1038/s41591-018-0272-7>
82. Wang J, Spicher N, Warnecke JM, Haghi M, Schwartze J, Deserno TM (2021) Unobtrusive health monitoring in private Spaces: the smart home. *Sensors* 21(3):864. <https://doi.org/10.3390/s21030864>
83. Kröger JL, Raschke P, Bhuiyan TR (2019) Privacy implications of accelerometer data: a review of possible inferences. In: Proceedings of the third international conference on cryptography, security and privacy – ICCSP ’19, Kuala Lumpur, Malaysia, pp 81–87. <https://doi.org/10.1145/3309074.3309076>
84. Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q* 13(3):319–340. <https://doi.org/10.2307/249008>
85. Moore K et al (2021) Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. *JMIR MHealth UHealth* 9(6):e23832. <https://doi.org/10.2196/23832>
86. Bagazzi RP (2007) The legacy of the technology acceptance model and a proposal for a paradigm shift. *J Assoc Inf Syst* 8(4):3
87. Venkatesh T, Xu X (2012) Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS Q* 36(1):157. <https://doi.org/10.2307/41410412>
88. Legris P, Ingham J, Collerette P (2003) Why do people use information technology? A critical review of the technology acceptance model. *Inf Manag* 40(3):191–204. [https://doi.org/10.1016/S0378-7206\(01\)00143-4](https://doi.org/10.1016/S0378-7206(01)00143-4)
89. Sarkar S, Chakrabarti D (2021) The perception and acceptance of wearable fitness devices among people and designing interventions for prolonged use. In: Ahram TZ, Falcão CS (eds) Advances in usability, user experience, wearable and assistive technology, vol 275. Springer International Publishing, Cham, pp 94–101. https://doi.org/10.1007/978-3-030-80091-8_12
90. Gemperle F, Kasabach C, Stivoric J, Bauer M, Martin R (1998) Design for wearability. In: Digest of Papers. Second international symposium on wearable computers (Cat. No. 98EX215), pp 116–122. <https://doi.org/10.1109/ISWC.1998.729537>
91. Zheng Y-L et al (2014) Unobtrusive Sensing and Wearable Devices for Health Informatics. *IEEE Trans Biomed Eng* 61(5):1538–1554. <https://doi.org/10.1109/TBME.2014.2309951>
92. European Commision (2014) Green paper on mobile health ('mHealth'). *Digit Agenda Eur*
93. Haut Autorité de santé (2019) Rapport d'analyse prospective 2019 Numérique: quelle (R)évolution?. [Online]. Available: https://www.has-sante.fr/upload/docs/application/pdf/2019-07/rapport_analyse_prospective_20191.pdf
94. U.S. Department of Health and Human Services Food and Drug Administration (2016) Use of real-world evidence to support regulatory decision-making for medical devices. Guidance for Industry and Food and Drug Administration Staff
95. Steinhubl SR, Muse ED, Topol EJ (2015) The emerging field of mobile health. *Sci Transl Med*

- 7(283):283rv3. <https://doi.org/10.1126/scitranslmed.aaa3487>
96. Torous J et al (2021) The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry* 20(3):318–335. <https://doi.org/10.1002/wps.20883>
97. Schwartz B (2004) The paradox of choice: Why more is less. HarperCollins Publishers, New York, p xi, 265
98. User engagement and abandonment of mHealth: a cross-sectional survey – PubMed. <https://pubmed.ncbi.nlm.nih.gov/35206837/> (accessed 07 Dec 2022)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part III

Methodologies



Chapter 13

Medical Image Segmentation Using Deep Learning

Han Liu, Dewei Hu, Hao Li, and Ipek Oguz

Abstract

Image segmentation plays an essential role in medical image analysis as it provides automated delineation of specific anatomical structures of interest and further enables many downstream tasks such as shape analysis and volume measurement. In particular, the rapid development of deep learning techniques in recent years has had a substantial impact in boosting the performance of segmentation algorithms by efficiently leveraging large amounts of labeled data to optimize complex models (supervised learning). However, the difficulty of obtaining manual labels for training can be a major obstacle for the implementation of learning-based methods for medical images. To address this problem, researchers have investigated many semi-supervised and unsupervised learning techniques to relax the labeling requirements. In this chapter, we present the basic ideas for deep learning-based segmentation as well as some current state-of-the-art approaches, organized by supervision type. Our goal is to provide the reader with some possible solutions for model selection, training strategies, and data manipulation given a specific segmentation task and dataset.

Key words Image segmentation, Deep learning, Semi-supervised method, Unsupervised method, Medical image analysis

1 Introduction

Image segmentation is an essential and challenging task in medical image analysis. Its goal is to delineate the object boundaries by assigning each pixel/voxel a label, where pixels/voxels with the same labels share similar properties or belong to the same class. In the context of neuroimaging, robust and accurate image segmentation can effectively help neurosurgeons and doctors, e.g., measure the size of brain lesions or quantitatively evaluate the volume changes of brain tissue throughout treatment or surgery. For instance, quantitative measurements of subcortical and cortical structures are critical for studies of several neurodegenerative diseases such as Alzheimer's, Parkinson's, and Huntington's diseases.

Authors Han Liu, Dewei Hu, and Hao Li have equal contributors to this chapter

Automatic segmentation of multiple sclerosis (MS) lesions is essential for the quantitative analysis of disease progression. The delineation of acute ischemic stroke lesions is crucial for increasing the likelihood of good clinical outcomes for the patient. While manual delineation of object boundaries is a tedious and time-consuming task, automatic segmentation algorithms can significantly reduce the workload of clinicians and increase the objectivity and reproducibility of measurements. To be specific, the segmentation task in medical images usually refers to semantic segmentation. For example, for paired brain structures (e.g., left and right pairs of subcortical structures), the instances of the same category will not be specified in the segmentation, in contrast to instance and panoptic segmentation.

There are many neuroimaging modalities such as magnetic resonance imaging, computed tomography, transcranial Doppler, and positron emission tomography. Moreover, neuroimaging studies often contain multimodal and/or longitudinal data, which can help improve our understanding of the anatomical and functional properties of the brain by utilizing complementary physical and physiological sensitivities. In this chapter, we first present some background information to help readers get familiar with the fundamental elements used in deep learning-based segmentation frameworks. Next, we discuss the learning-based segmentation approaches in the context of different supervision settings, along with some real-world applications.

2 Methods

2.1 Fundamentals

2.1.1 Common Network Architectures for Segmentation Tasks

Convolutional neural networks (CNNs) dominated the medical image segmentation field in recent years. CNNs leverage information from images to predict segmentations by hierarchically learning parameters with linear and nonlinear layers. We begin by discussing some popular models and their architectures: (1) U-Net [1], (2) V-Net [2], (3) attention U-Net [3, 4], and (4) nnU-Net [5, 6].

U-Net is the most popular model for medical image segmentation, and its architecture is shown in Fig. 1. The network has two main parts: the encoder and the decoder, with skip connections in between. The encoder consists of two repeated 3×3 convolutions (conv) without zero-padding, a rectified linear unit (ReLU) activation function. A max-pooling operation with stride 2 is used for connecting different levels or downsampling. We note that the channel number of feature maps is doubled at each subsequent level. In the symmetric decoder counterpart, a 2×2 up-convolution (up-conv) is used not only for upsampling but also for reducing the number of channels by half. The center-cropped feature map from the encoder is delivered to the decoder

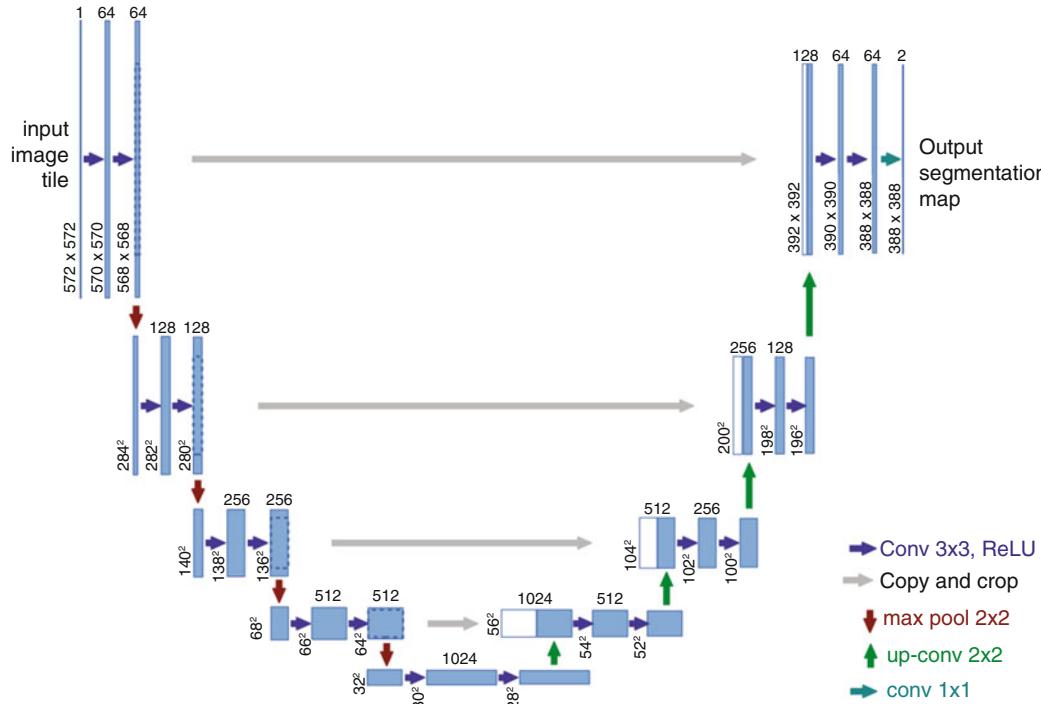


Fig. 1 U-Net architecture. Blue boxes are the feature maps. Channel numbers are denoted above each box, while the tensor sizes are denoted on the lower left. White boxes show the concatenations and arrows indicate various operations. ©2015 Springer Nature. Reprinted, with permission, from [1]

via skip connections at each level to preserve the low-level information. The cropping is needed to maintain the same size between feature maps for concatenation. Next, two repeated 3×3 conv and ReLU are applied. Lastly, a 1×1 conv is employed for converting the channel number to the desired number of classes C . In this configuration, the network takes a 2D image as input and produces a segmentation map with C classes. Later, a 3D U-Net [7] was introduced for volumetric segmentation that learns from volumetric images.

V-Net is another popular model for volumetric medical image segmentation. Based upon the overall structure of the U-Net, the V-Net [2] leverages the residual block [8] to replace the regular conv, and the convolution kernel size is enlarged to $5 \times 5 \times 5$. The residual blocks can be formulated as follows: (1) the input of a residual block is processed by conv layers and nonlinearities, and (2) the input is added to the output from the last conv layer or nonlinearity of the residual block. It consists of a fully convolutional neural network trained end-to-end.

Attention U-Net is a model based on U-Net with attention gates (AG) in the skip connections (Fig. 2). The attention gates can learn to focus on the segmentation target. The salient features are

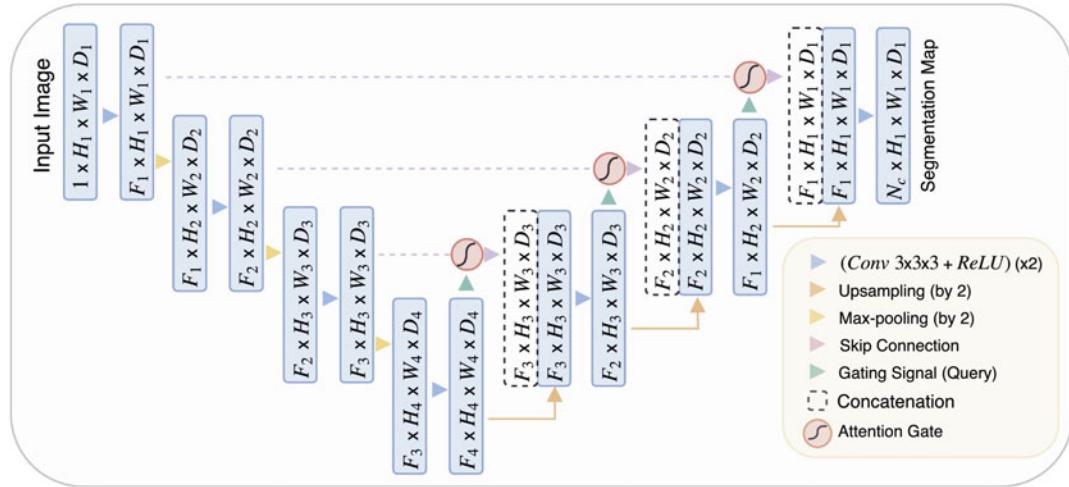


Fig. 2 Attention U-Net architecture. H_i , W_i , and D_i represent the height, width, and depth of the feature map at the i^{th} layer of the U-Net structure. F_i indicates the number of feature map channels. Replicated from [4] (CC BY 4.0)

emphasized with larger weights from the CNN during the training. This leads the model to achieve higher accuracy on target structures with various shapes and sizes. In addition, AGs are easy to integrate into the existing popular CNN architectures. The details of the attention mechanism and attention gates are discussed in Subheading 2.1.2. More details on attention can also be found in Chap. 6.

nnU-Net is a medical image segmentation pipeline that can achieve a self-configuring network architecture based on the different datasets and tasks it is given, without any manual intervention. According to the dataset and task, nnU-Net will generate one of (1) 2D U-Net, (2) 3D U-Net, and (3) cascaded 3D U-Net for the segmentation network. For cascaded 3D U-Net, the first network takes downsampled images as inputs, and the second network uses the image at full resolution as input to refine the segmentation accuracy. The nnU-Net is often used as a baseline method in many medical image segmentation challenges, because of its robust performance across various target structures and image properties. The details of nnU-Net can be found in [6].

2.1.2 Attention Modules

Although the U-Net architecture described in Subheading 2.1.1 has achieved remarkable success in medical image segmentation, the downsampling steps included in the encoder path can induce poor segmentation accuracy for small-scale anatomical structures (e.g., tumors and lesions). To tackle this issue, the attention modules are often applied so that the salient features are enhanced by higher weights, while the less important features are ignored. This subsection will introduce two types of attention mechanisms: additive attention and multiplicative attention.

Additive Attention As discussed in the previous section, U-Net is the most popular backbone for medical image analysis tasks. The downsampling enables it to work on features of different scales. Suppose we are working on a 3D segmentation problem. The output of the U-Net encoder at the l th level is then a tensor X^l of size $[F_l, H_l, W_l, D_l]$, where H_l, W_l, D_l denote the height, width, and depth of the feature map, respectively, and F_l represents the length of the feature vectors. We regard the tensor as a set of feature vectors \mathbf{x}_i^l :

$$\mathcal{X}^l = \{\mathbf{x}_i^l\}_{i=1}^n, \quad \mathbf{x}_i^l \in \mathbb{R}^{F_l} \quad (1)$$

where $n = H_l \times W_l \times D_l$. The attention gate assigns a weight α_i to each vector \mathbf{x}_i so that the model can concentrate on salient features. Ideally, important features are assigned higher weight that will not vanish when downsampling. The output of the attention gate will be a collection of weighted feature vectors:

$$\hat{\mathcal{X}}^l = \{\alpha_i^l \cdot \mathbf{x}_i^l\}_{i=1}^n, \quad \alpha_i^l \in \mathbb{R} \quad (2)$$

These weights α_i , also known as gating coefficients, are determined by an attention mechanism that delineates the correlation between the feature vector \mathbf{x} and a gating signal \mathbf{g} . As shown in Fig. 3, for all $\mathbf{x}_i^l \in \mathcal{X}^l$, we compute an additive attention with regard to a corresponding \mathbf{g}_i by

$$s_{att}^l = \psi^\top \left[\sigma_1 \left(\mathbf{W}_x^\top \mathbf{x}_i^l + \mathbf{W}_g^\top \mathbf{g}_i + b_g \right) \right] + b_\psi \quad (3)$$

where b_g and b_ψ represent the bias and $\mathbf{W}_x, \mathbf{W}_g, \psi$ are linear transformations. The output dimension of the linear transformation is $\mathbb{R}^{F_{int}}$ where F_{int} is a self-defined integer. Denote these

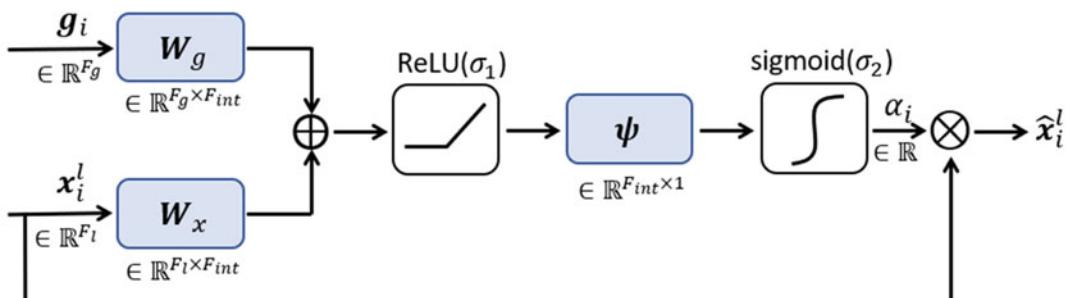


Fig. 3 The structure of the additive attention gate. x_i^l is the i th feature vector at the l th level of the U-Net structure and \mathbf{g}_i is the corresponding gating signal. \mathbf{W}_x and \mathbf{W}_g are the linear transformation matrices applied to x_i^l and \mathbf{g}_i , respectively. The sum of the resultant vectors will be activated by ReLU and then its dot product with a vector ψ is computed. The sigmoid function is used to normalize the resulting scalar to $[0, 1]$ range, which is the gating coefficient α_i . The weighted feature vector is denoted by \hat{x}_i^l . Adapted from [4] (CC BY 4.0)

learnable parameters by a set Θ_{att} . The coefficients s_{att}^l are normalized to $[0, 1]$ by a sigmoid function σ_2 :

$$\alpha_i^l = \sigma_2(s_{att}^l(\mathbf{x}_i^l, \mathbf{g}_i; \Theta_{att})) \quad (4)$$

Basically, the attention gate is thus a linear combination of the feature vector and the gating signal. In practical applications [3, 4, 9], the gating signal is chosen to be the coarser feature space as indicated in Fig. 2. In other words, for input feature \mathbf{x}_i^l , the corresponding gating signal is defined by

$$\mathbf{g}_i = \mathbf{x}_i^{l+1} \quad (5)$$

Note that an extra downsampling step should be applied on \mathbf{X}^l so that it has the same shape as \mathbf{X}^{l+1} . In experiments to segment brain tumor on MRI datasets [9] and the pancreas on CT abdominal datasets [4], AG was shown to improve the segmentation performance for diverse types of model backbones including U-Net and Residual U-Net.

Multiplicative Attention Similar to additive attention, the multiplicative mechanism can also be leveraged to compute the importance of feature vectors. The basic idea of multiplicative attention was first introduced in machine translation [11]. Evolving from that, Vaswani et al. proposed a groundbreaking transformer architecture [10] which has been widely implemented in image processing [12, 13]. In recent research, transformers have been incorporated with the U-Net structure [14, 15] to improve medical image segmentation performance.

The attention function is described by matching a query vector \mathbf{q} with a set of key vectors $\{\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_n\}$ to obtain the weights of the corresponding values $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$. Figure 4a shows an example for $n=4$. Suppose the vectors \mathbf{q} , \mathbf{k}_i , and \mathbf{v}_i have the same dimension \mathbb{R}^d . Then, the attention function is

$$s_i = \frac{\mathbf{q}^\top \mathbf{k}_i}{\sqrt{d}} \quad (6)$$

We note that the dot product can have large magnitude when d is large, which can cause gradient vanishing problem in the softmax function; s_i is normalized by the size of the vector to alleviate this. Equation 13.6 is a commonly used attention function in transformers. There are some other options including $s_i = \mathbf{q}^\top \mathbf{k}_i$ and $s_i = \mathbf{q}^\top \mathbf{Wk}_i$ where \mathbf{W} is a learnable parameter. Generally, the attention value s_i is determined by the similarity between the query and the key. Similar to the additive attention gate, these attention values are normalized to $[0, 1]$ by a softmax function σ_3 :

$$\alpha_i = \sigma_3(s_1, \dots, s_n) = \frac{e^{s_i}}{\sum_{j=1}^n e^{s_j}} \quad (7)$$

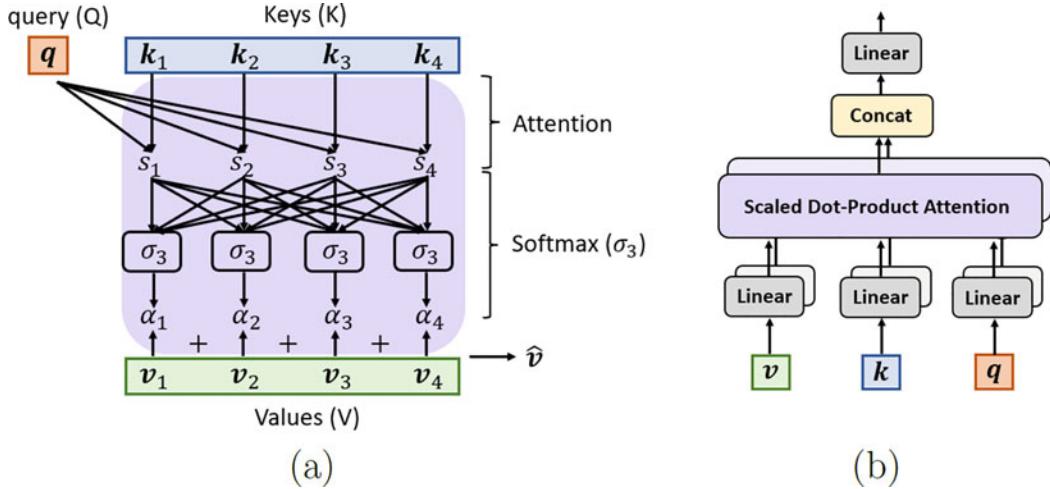


Fig. 4 (a) The dot-product attention gate. k_i are the keys and q is the query vector. s_i are the outputs of the attention function. By using the softmax σ_3 , the attention coefficients α_i are normalized to [0, 1] range. The output will be the weighted sum of values v_i . (b) The multi-head attention is implemented in transformers. The input values, keys, and query are linearly projected to different spaces. Then the dot-product attention is applied on each space. The resultant vectors are concatenated by channel and passed through another linear transformation. Image (b) is adapted from [10]. Permission to reuse was kindly granted by the authors

The output of the attention gate will be $\hat{v} = \sum_{i=1}^n \alpha_i v_i$. In the transformer application, the values, keys, and queries are usually linearly projected into several different spaces, and then the attention gate is applied in each space as illustrated in Fig. 4b. This approach is called multi-head attention; it enables the model to jointly attend to information from different subspaces.

In practice, the value v_i is often defined by the same feature vector as the key k_i . This is why the module is also called multi-head self-attention (MSA). Chen et al. proposed the TransUNet [15], which leverages this module in the bottleneck of a U-Net as shown in Fig. 5. They argue that such a combination of a U-Net and the transformer achieves superior performance in multi-organ segmentation tasks.

2.1.3 Loss Functions for Segmentation Tasks

This section summarizes some of the most widely used loss functions for medical image segmentation (Fig. 6) and describes their usage in different scenarios. A complementary reading material for an extensive list of loss functions can be found in [16, 17]. In the following, the predicted probability by the segmentation model and the ground truth at the i th pixel/voxel are denoted as p_i and g_i , respectively. N is the number of voxels in the image.

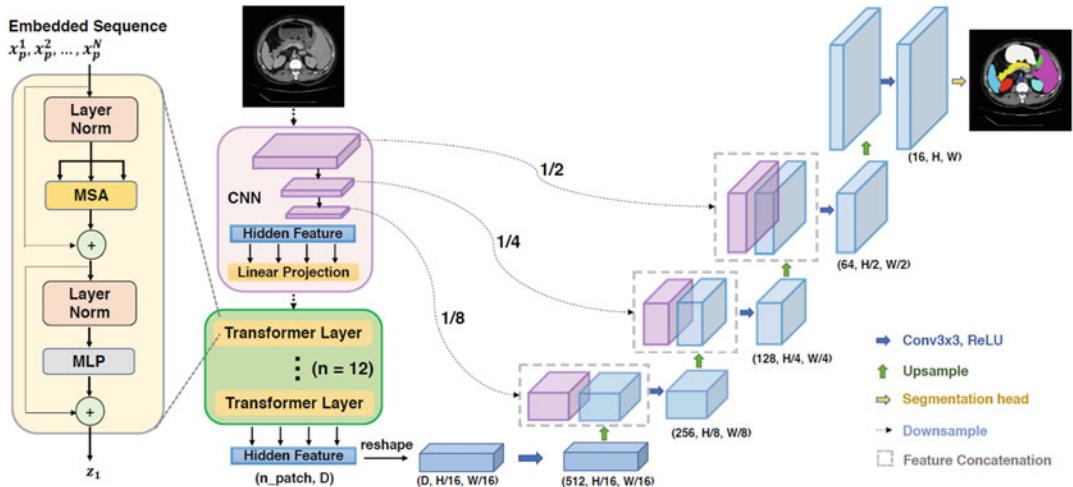


Fig. 5 The architecture of TransUNet. The transformer layer represented by the yellow box shows the application of multi-head attention (MSA). MLP represents the multilayer perceptron. In general, the feature vectors in the bottleneck of the U-Net are set as the input to the stack of n transformer layers. As these layers will not change the dimension of the features, they are easy to be implemented and will not affect other parts of the U-Net model. Replicated from [15] (CC BY 4.0)

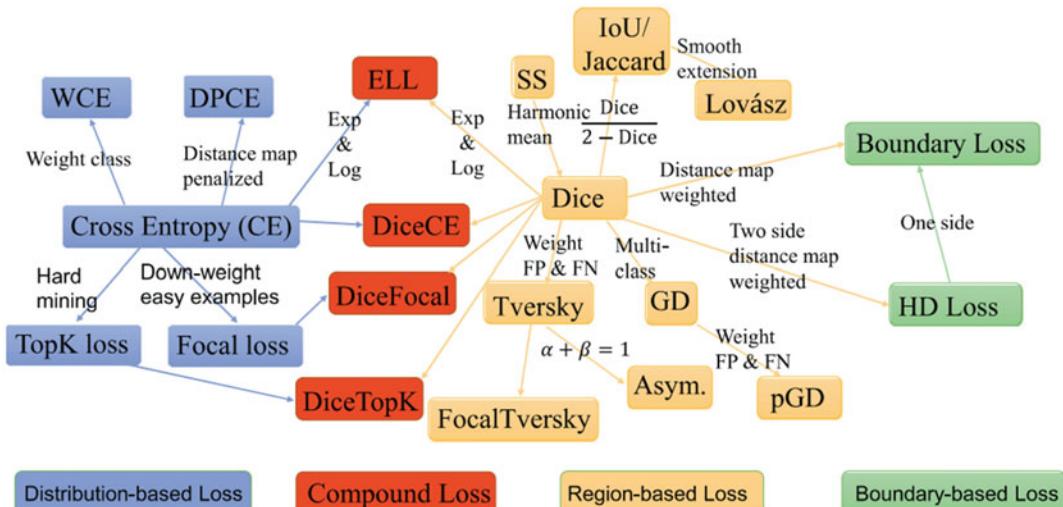


Fig. 6 Loss functions for medical image segmentation. WCE: weighted cross-entropy loss. DPCE: distance map penalized cross-entropy loss. ELL: exponential logarithmic loss. SS: sensitivity-specificity loss. GD: generalized Dice loss. pGD: penalty loss. Asym: asymmetric similarity loss. IoU: intersection over union loss. HD: Hausdorff distance loss. ©2021 Elsevier. Reprinted, with permission, from [16]

Cross-Entropy Loss Cross-entropy (CE) is defined as a measure of the difference between two probability distributions for a given random variable or set of events. This loss function is used for pixel-wise classification in segmentation tasks:

$$\ell_{CE} = - \sum_i^N \sum_k^K y_i^k \log(p_i^k) \quad (8)$$

where N is the number of voxels, K is the number of classes, y_i^k is a binary indicator that shows whether k is the correct class, and p_i^k is the predicted probability for voxel i to be in k th class.

Weighted Cross-Entropy Loss Weighted cross-entropy (WCE) loss is a variant of the cross-entropy loss to address the class imbalance issue. Specifically, class-specific coefficients are used to weigh each class differently, as follows:

$$\ell_{WCE} = - \sum_i^N \sum_k^K w_{y_k} y_i^k \log(p_i^k) \quad (9)$$

Here, w_{y_k} is the coefficient for the k th class. Suppose there are 5 positive samples and 12 negative samples in a binary classification training set. By setting $w_0=1$ and $w_1=2$, the loss would be as if there were ten positive samples.

Focal Loss Focal loss was proposed to apply a modulating term to the CE loss to focus on hard negative samples. It is a dynamically scaled CE loss, where the scaling factor decays to zero as confidence in the correct class increases. Intuitively, this scaling factor can automatically down-weight the contribution of easy examples during training and rapidly focus the model on hard examples:

$$\ell_{Focal} = - \sum_i^N \alpha_i (1 - p_i)^\gamma \log(p_i) \quad (10)$$

Here, α_i is the weighing factor to address the class imbalance and γ is a tunable focusing parameter ($\gamma > 0$).

Dice Loss The Dice coefficient is a widely used metric in the computer vision community to calculate the similarity between two binary segmentations. In 2016, this metric was adapted as a loss function for 3D medical image segmentation [2]:

$$\ell_{Dice} = 1 - \frac{2 \sum_i^N p_i g_i + 1}{\sum_i^N (p_i + g_i) + 1} \quad (11)$$

Generalized Dice Loss Generalized Dice loss (GDL) [18] was proposed to reduce the well-known correlation between region size and Dice score:

$$L_{GDL} = 1 - 2 \frac{\sum_{l=1}^2 w_l \sum_i^N p_i g_i}{\sum_{l=1}^2 w_l \sum_i^N p_i + g_i} \quad (12)$$

Here $w_l = \frac{1}{(\sum_i^N g_i)^2}$ is used to provide invariance to different region sizes, i.e., the contribution of each region is corrected by the inverse of its volume.

Tversky Loss The Tversky loss [19] is a generalization of the Dice loss by adding two weighting factors α and β to the FP (false positive) and FN (false negative) terms. The Tversky loss is defined as

$$L_{Tversky} = 1 - \frac{\sum_i^N p_i g_i}{\sum_i^N p_i g_i + \alpha(1-g_i)p_i + \beta(1-p_i)g_i} \quad (13)$$

Recently, a comprehensive study [16] of loss functions on medical image segmentation tasks shows that using Dice-related compound loss functions, e.g., Dice loss + CE loss, is a better choice for new segmentation tasks, though none of losses can consistently achieve the best performance on multiple segmentation tasks. Therefore, for a new segmentation task, we recommend the readers to start with Dice + CE loss, which is also the default loss function in one of the most popular medical image segmentation frameworks, nnU-Net [6].

Finally, note that other loss functions have also been proposed to introduce prior knowledge about size, topology, or shape, for instance [20].

2.1.4 Early Stopping

Given a loss function, a simple strategy for training is to stop the training process once a predetermined maximum number of iterations are reached. However, too few iterations would lead to an under-fitting problem, while over-fitting may occur with too many iterations. “Early stopping” is a potential method to avoid such issues. The training set is split into training and validation sets when using the early stopping condition. The early stopping condition is based on the performance on the validation set. For example, if the validation performance (e.g., average Dice score) does not increase for a number of iterations, the early stopping condition is triggered. In this situation, the best model with the highest performance on the validation set is saved and used for inference. Of course, one should not report the validation performance for the validation of the model. Instead, one should use a separate test set which is kept unseen during training for an unbiased evaluation.

2.1.5 Evaluation Metrics for Segmentation Tasks

Various metrics can quantitatively evaluate different aspects of a segmentation algorithm. In a binary segmentation task, a true positive (TP) indicates that a pixel in the target object is correctly predicted as target. Similarly, a true negative (TN) represents a background pixel that is correctly identified as background. On the other hand, a false positive (FP) and a false negative (FN) refer to a wrong prediction for pixels in the target and background, respectively. Most of the evaluation metrics are based upon the number of pixels in these four categories.

Sensitivity measures the completeness of positive predictions with regard to the positive ground truth (TP + FN). It thus shows the model's ability to identify target pixels. It is also referred to as recall or true-positive rate (TPR). It is defined as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (14)$$

As the negative counterpart of sensitivity, specificity describes the proportion of negative pixels that are correctly predicted. It is also referred to as true-negative rate (TNR). It is defined as

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (15)$$

Specificity can be difficult to interpret because TN is usually very large. It can even be misleading as TN can be made arbitrarily large by changing the field of view. This is due to the fact that the metric is computed over pixels and not over patients/controls like in classification tasks (the number of controls is fixed). In order to provide meaningful measures of specificity, it is preferable to define a background region that has an anatomical definition (for instance, the brain mask from which the target is subtracted) and does not include the full field of view of the image.

Positive predictive value (PPV), also known as precision, measures the correct rate among pixels that are predicted as positives:

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (16)$$

For clinical interpretation of segmentation, it is often useful to have a more direct estimation of false negatives. To that purpose, one can report the false discovery rate:

$$\text{FDR} = 1 - \text{PPV} = \frac{\text{FP}}{\text{TP} + \text{FP}} \quad (17)$$

which is redundant with PPV but may be more intuitive for clinicians in the context of segmentation.

Dice similarity coefficient (DSC) measures the proportion of spatial overlap between the ground truth (TP+FN) and the predicted positives (TP+FP). Dice similarity is the same as the F_1 score, which computes the harmonic mean of sensitivity and PPV:

$$\text{DSC} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \quad (18)$$

Accuracy is the ratio of correct predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (19)$$

As was the case in specificity, we note that there are many segmentation tasks where the target anatomical structure is very small (e.g., subcortical structures); hence, the foreground and background have unbalanced number of pixels. In this case, accuracy can be misleading and display high values for poor segmentations. Moreover, as for the case of specificity, one needs to define a background region in order for TN, and thus accuracy, not to vary arbitrarily with the field of view.

The Jaccard index (JI), also known as the intersection over union (IoU), measures the percentage of overlap between the ground truth and positive prediction relative to the union of the two:

$$\text{JI} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (20)$$

JI is closely related to the DSC. However, it is always lower than the DSC and tends to penalize more severely poor segmentations.

There are also distance measures of segmentation accuracy which are especially relevant when the accuracy of the boundary is critical. These include the average symmetric surface distance (ASSD) and the Hausdorff distance (HD). Suppose the surface of the ground truth and the predicted segmentation are \mathcal{S} and \mathcal{S}' , respectively. For any point $\mathbf{p} \in \mathcal{S}$, the distance from \mathbf{p} to surface \mathcal{S}' is defined by the minimum Euclidean distance:

$$d(\mathbf{p}, \mathcal{S}') = \min_{\mathbf{p}' \in \mathcal{S}'} \|\mathbf{p} - \mathbf{p}'\|_2 \quad (21)$$

Then the average distance between \mathcal{S} and \mathcal{S}' is given by averaging over \mathcal{S} :

$$d(\mathcal{S}, \mathcal{S}') = \frac{1}{N_S} \sum_{i=1}^{N_S} d(\mathbf{p}_i, \mathcal{S}') \quad (22)$$

Note that $d(\mathcal{S}, \mathcal{S}') \neq d(\mathcal{S}', \mathcal{S})$. Therefore, both directions are included in ASSD so that the mean of the surface distance is symmetric:

$$\text{ASSD} = \frac{1}{N_S + N_{S'}} \left[\sum_{i=1}^{N_S} d(\mathbf{p}_i, \mathcal{S}') + \sum_{j=1}^{N_{S'}} d(\mathbf{p}'_j, \mathcal{S}) \right] \quad (23)$$

The ASSD tends to obscure localized errors when the segmentation is decent at most of the points on the boundary. The Hausdorff distance (HD) can better represent the error, by, instead of

computing the average distance to a surface, computing the maximum distance. To that purpose, one defines

$$h(S, S') = \max_{p \in S} d(p, S') \quad (24)$$

Note that, again, $h(S, S') \neq h(S', S)$. Therefore, both directions are included in HD so that the distance is symmetric:

$$HD = \max(h(S, S'), h(S', S)) \quad (25)$$

HD is more sensitive than ASSD to localized errors. However, it can be too sensitive to outliers. Hence, using the 95th percentile rather than the maximum value for computing $h(S, S')$ is a good option to alleviate the problem.

Moreover, there are some volume-based measurements that focus on correctly estimating the volume of the target structure, which is essential for clinicians since the size of the tissue is an important marker in many diseases. Denote the ground truth volume as V while the prediction volume as V' . There are a few expressions for the volume difference. (1) The unsigned volume difference: $|V' - V|$. (2) The normalized unsigned difference: $\frac{|V' - V|}{V}$. (3) The normalized signed difference: $\frac{V' - V}{V}$. (4) Pearson's correlation coefficient between the ground truth volumes and the predicted volumes: $\frac{\text{Cov}(V, V')}{\sqrt{\text{Var}(V)}\sqrt{\text{Var}(V')}}$. Nevertheless, note that, while they are useful, these volume-based metrics can also be misleading (a segmentation could be wrongly placed while providing a reasonable volume estimate) when used in isolation. They thus need to be combined with overlap metrics such as Dice.

Finally, some recent guidelines on validation of different image analysis tasks, including segmentation, were published in [21].

2.1.6 Pre-processing for Segmentation Tasks

Image pre-processing is a set of sequential steps taken to improve the data and prepare it for subsequent analysis. Appropriate image pre-processing steps often significantly improve the quality of feature extraction and the downstream image analysis. For deep learning methods, they can also help the training process converge faster and achieve better model performance. The following sections will discuss some of the most widely used image pre-processing techniques.

Skull Stripping Many neuroimaging applications often require preliminary processing to isolate the brain from extracranial or non-brain tissues from MRI scans, commonly referred to as skull stripping. Skull stripping helps reduce the variability in datasets and is a critical step prior to many other image processing algorithms such as registration, segmentation, or cortical surface reconstruction. In literature, skull stripping methods are broadly classified into five categories: mathematical morphology-based methods

[22], intensity-based methods [23], deformable surface-based methods [24], atlas-based methods [25], and hybrid methods [26]. Recently, deep learning-based skull stripping methods have been proposed [27–32] to improve the accuracy and efficiency. A detailed discussion of the merits and limitations of various skull stripping techniques can be found in [33].

Bias Field Correction The bias field refers to a low-frequency and very smooth signal that corrupts MR images [34]. These artifacts, often described as shading or bias, can be generated by imperfections in the field coils or by magnetic susceptibility changes at the boundaries between anatomical tissue and air. This bias field can significantly degrade the performance of image processing algorithms that use the image intensity values. Therefore, a pre-processing step is usually required to remove the bias field. The N4 bias field correction algorithm [35] is one of the most widely used methods for this purpose, as it assumes a simple parametric model and does not require tissue classification.

Data Harmonization Another challenge of MRI data is that it suffers from significant intensity variability due to several factors such as variations in hardware, reconstruction algorithms, and acquisition settings. This is also due to the fact that most MR imaging sequences (e.g., T1-weighted, T2-weighted) are not quantitative (the voxel values can only be interpreted relative to each other). Such differences can often be pronounced in multisite studies, among others. This variability can be problematic because intensity-based models may not generalize well to such heterogeneous datasets. Any resulting data can suffer from significant biases caused by acquisition details rather than anatomical differences. It is thus desirable to have robust data harmonization methods to reduce unwanted variability across sites, scanners, and acquisition protocols. One of the popular MRI harmonization methods is a statistical approach named the combined association test (comBat). This method was shown to exhibit a good capacity to remove unwanted site biases while preserving the desired biological information [36]. Another popular method is a deep learning-based image-to-image translation model, CycleGAN [37]. The CycleGAN and its variants do not require paired data, and thus the training process is unsupervised in the context of data harmonization.

Intensity Normalization Intensity normalization is another important step to ensure comparability across images. In this section, we discuss common intensity normalization techniques. Readers can refer to the work [38] in which the author explores the impact of different intensity normalization techniques on MR image synthesis.

Z-Score Normalization The basic Z-score normalization on the entire image is also called the whole-brain normalization. Given the mean μ and standard deviation σ from all voxels in a brain mask B , Z-score normalization can be performed for all voxels in image I as follows:

$$I_{z-score}(x) = \frac{I(x) - \mu}{\sigma} \quad (26)$$

While straightforward to implement, whole-brain normalization is known to be sensitive to outliers.

White Stripe Normalization White stripe normalization [39] is based on the parameters obtained from a sample of normal-appearing white matter (NAWM) and is thus robust to local intensity outliers such as lesions. The NAWM is obtained by smoothing the histogram of the image I and selecting the mode of the distribution. For T1-weighted MRI, the “white stripe” is defined as the 10% of intensity values around the mean of NAWM μ . Let $F(x)$ be the CDF of the specific MR image $I(x)$ inside the brain mask B , and $\tau = 5\%$. The white stripe Ω_τ is defined as

$$\Omega_\tau = \{I(x) | F^{-1}(F(x) - \tau) < I(x) < F^{-1}(F(x) + \tau)\} \quad (27)$$

Then let σ_τ be the sample standard deviation associated with Ω_τ . The white stripe normalized image is

$$I_{ws}(x) = \frac{I(x) - \mu}{\sigma_\tau} \quad (28)$$

Compared to the whole-brain normalization, the white stripe normalization may work better and have better interpretation, especially for applications where intensity outliers such as lesions are expected.

Segmentation-Based Normalization Segmentation-based normalization uses a segmentation of a specified tissue, such as the cerebrospinal fluid (CSF), gray matter (GM), or white matter (WM), to normalize the entire image to the mean of the tissue. Let $T \subset B$ be the tissue mask for image I . The tissue mean can be calculated as $\mu = \frac{1}{|T|} \sum_{t \in T} I(t)$ and the segmentation-based normalized image is expressed as

$$I_{seg}(x) = \frac{cI(x)}{\mu} \quad (29)$$

where $c \in \mathbb{R}^+$ is a constant.

Kernel Density Estimate Normalization Kernel density estimate (KDE) normalization estimates the empirical probability density function of the intensities of the entire image I over the brain

mask B via kernel density estimation. The KDE of the probability density function for the image intensities can be expressed as

$$\hat{p}(x) = \frac{1}{HWD \times \delta} \sum_{i=1}^{HWD} K\left(\frac{x - x_i}{\delta}\right) \quad (30)$$

where H , W , D are the image sizes of I , x is an intensity value, K is the kernel, and δ is the bandwidth parameter which scales the kernel. With KDE normalization, the mode of WM can be selected more robustly via a smooth version of the histogram and thus is more suitable to be used in a segmentation-based normalization method.

Spatial Normalization Spatial normalization aims to register a subject's brain image to a common space (reference space) to allow comparisons across subjects. When the reference space is a standard space, such as the Montreal Neurological Institute (MNI) space [40] or the Talairach and Tournoux atlas (Talairach space), the registration also facilitates the sharing and interpretation of data across studies. It is also common practice to define a customized space from a dataset rather than using a standard space. For deep learning methods, it has been shown that training data with appropriate spatial normalization tend to yield better performances [41–43]. Rigid, affine, or deformable registration may be desirable for spatial normalization, depending on the application. Many registration methods are publicly available through software packages such as 3D Slicer, FreeSurfer [<https://surfer.nmr.mgh.harvard.edu/>], FMRIB Software Library (FSL) [<https://fsl.fmrib.ox.ac.uk/fsl/fslwiki>], and Advanced Normalization Tools (ANTs) [<https://picsl.upenn.edu/software/ants/>].

2.2 Supervision Settings

In the following three sections, we categorize the learning-based segmentation algorithms by their supervision setting. In the reverse order of the amount of annotation required, these include supervised, semi-supervised, and unsupervised methods (Fig. 7). For supervised methods, we mainly present some training strategies and model architectures that will help improve the segmentation performance. For the other two types of approaches, we classify the mainstream ideas and then provide application examples proposed in recent research.

2.3 Supervised Methods

2.3.1 Background

In supervised learning, a model is presented with the given dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$ of inputs x and associated labels y . This y can take several forms, depending on the learning task. In particular, for fully convolutional neural network-based segmentation applications, y is a segmentation map. In supervised learning, the model can learn from labeled training data by minimizing the loss function

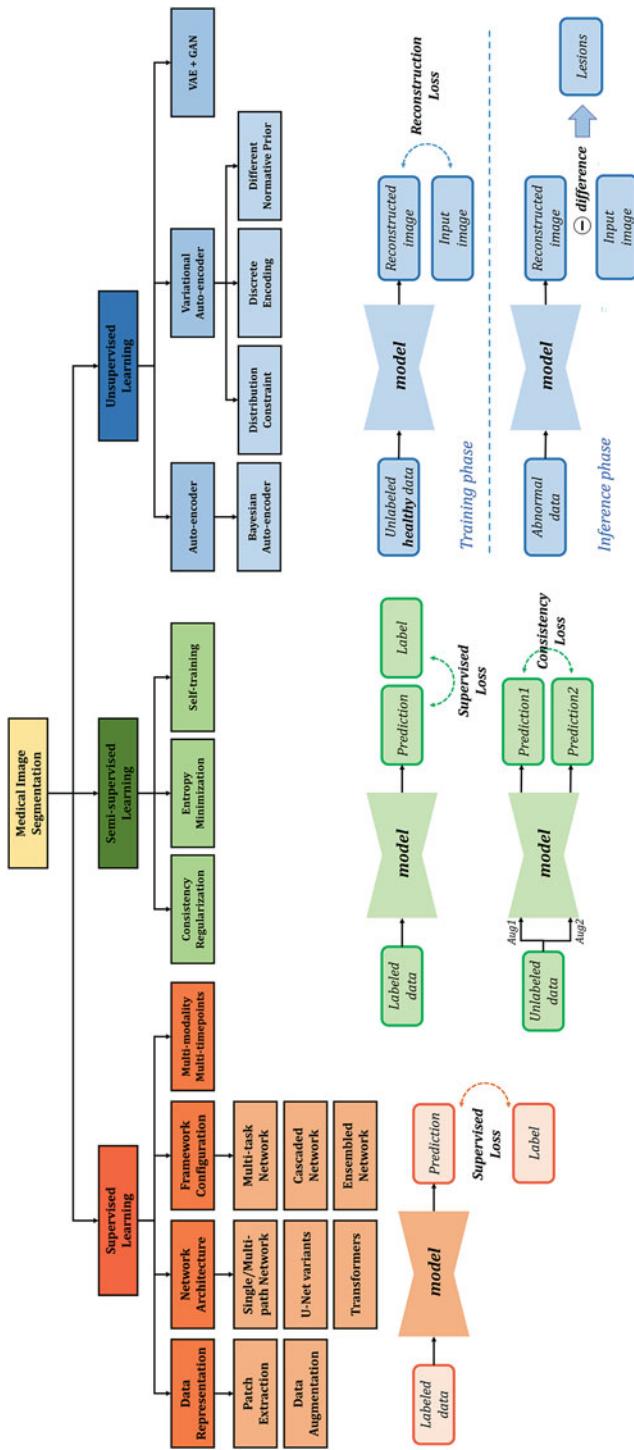


Fig. 7 Overview of the supervision settings for medical image segmentation. Best viewed in color

and apply what it has learned to make a prediction/segmentation in testing data. Supervised training thus aims to find model parameters θ that best predict the data based on a loss function $L(y, \hat{y})$. Here, \hat{y} denotes the output of the model obtained by feeding a data point x to the function $f(x; \theta)$ that represents the model. Given sufficient training data, supervised methods can generally perform better than semi-supervised or unsupervised segmentation methods.

2.3.2 Data Representation

Data is an important part of supervised segmentation models, and the model performance relies on data representation. In addition to image pre-processing (Subheading 2.1.6), there are a few key steps for data preparation before being fed into the segmentation network.

Patch Formulation The inputs of CNN can be represented as image patches when the whole image is too large and would require too much GPU memory. The image patches could be 2D slices, 3D patches, and any format in between. The choice of patches would affect the performance of networks for a given dataset and task [44]. Compared to 3D patches, 2D slices have the advantage of lighter computational load during training. However, contextual information along the third axis is missing. In contrast, 3D patches leverage data from all three axes, but they require more computational resources. As a compromise between 2D and 3D patches, “2.5D” approaches have been proposed, by taking 2D slices in all three orthogonal views through the same voxel [45]. Those 2D slices could be trained in a single CNN or a separate CNN for each view. Furthermore, Zhang et al. [46] proposed 2.5D stacked slices to leverage the information from adjacent slices in each view.

Patch Extraction Due to the imbalance between foreground and background, various patch extraction strategies have been designed to obtain robust segmentation. Kamnitsas et al. [47], Dolz et al. [48], and Li et al. [49] pick a voxel within the foreground or background with 50% probability at every iteration during training and select the patch centered at that voxel. In [46], Zhang et al. extract 2.5D stacked patches if the central slice contains the foreground, even with only one voxel. In some models [50, 51], 3D patches with target structure are used as input instead of the whole image, which could reduce the effect of the background for segmenting target structures with smaller volume.

Data Augmentation To avoid the over-fitting problem and increase the generalizability of the model, data augmentation (DA) is widely used in medical image segmentation [52]. The common DA strategies could be classified into three categories:

(1) spatial augmentation, (2) image appearance augmentation, and (3) image quality augmentation. For spatial augmentation, random image flip, rotation, scale, and deformation are often used [4, 45, 53–55]. Random gamma correction, intensity scale, and intensity shift are the common forms for image appearance augmentation [51, 54, 56, 57]. Image quality augmentation includes random Gaussian blur, random noise addition, and image sharpening [51, 56]. Note that while we only list a few commonly used methods here, many others have been explored. TorchIO [58] is a widely used software package for data augmentation.

2.3.3 Network Architecture

Here, we classify the popular supervised segmentation networks into single/multipath networks and encoder-decoder networks.

Single/Multipath Networks As discussed above, patches are often used as input instead of the entire image, resulting in a lack of global context. This could produce noisy segmentations, such as undesired islands of false-positive voxels that need to be removed in post-processing [48]. To compensate for the missing global context, Li et al. [49] used spatial coordinates as additional channels of input patches. A multipath network is another feasible solution (Fig. 8). Multipath networks usually contain global and local paths [47, 59, 60] that extract different features at different scales. The global path uses convolutions with larger kernel size [60] or a larger receptive field [47] to learn global information [47]. In

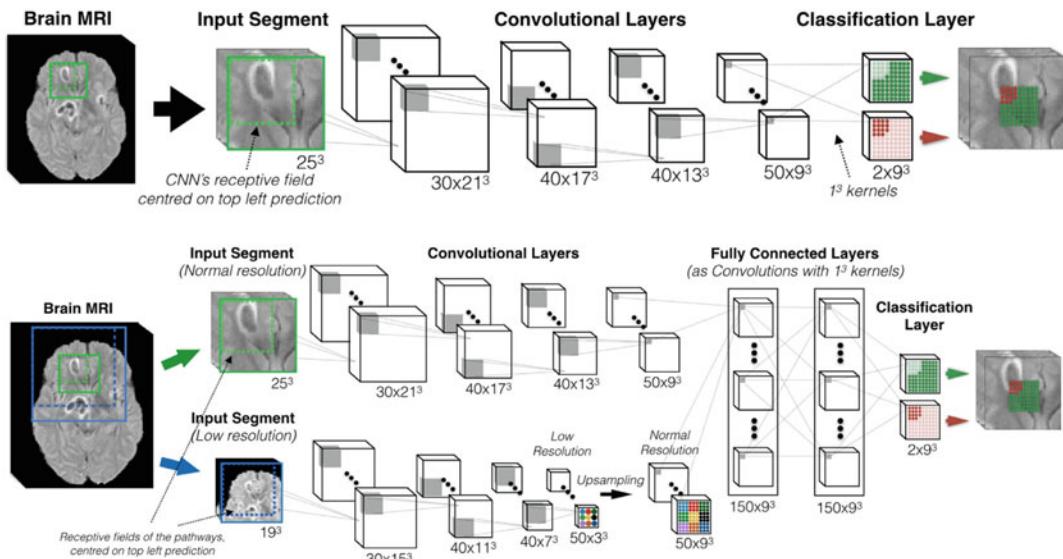


Fig. 8 Examples of single-path (top) and multipath (bottom) networks. In the multipath network, the inputs for the two pathways are centered at the same location. The top pathway is equivalent to the single-path network and takes the normal resolution image as input, while the bottom pathway takes a downsampled image with larger field of view as input. Replicated from [47] (CC BY 4.0)

contrast, local features are extracted in the local path. The global path thus extracts global features and tends to locate the position of the target structure. In contrast, the shape, size, texture, boundary, and other details of the target structure are identified by the local path. However, the performance of this type of network is easily affected by the size and design of input patches: for example, too small patches would not provide enough information, while too large patches would be computationally prohibitive.

U-Net and Its Variants To tackle the limitations of the single/multipath networks, many models use U-net variants with encoder-decoder paths [1, 61], which establishes end-to-end training from image to segmentation map. The encoder is similar to the single/multipath networks but with downsampling operations between the different scales of feature maps. The decoder leverages the extracted features from the encoder and produces a segmentation of the same size as the original image. Skip connections that pass the feature maps from the encoder directly to the decoder contribute to the performance of the U-net. The passed information could help to recover the details of segmentation.

The most common modification of the U-Net is the introduction of other convolutional modules, such as *residual blocks* [62], *dense blocks* [63], *attention modules* [3, 4], etc. These convolutional modules could replace regular convolution operations or be used in the skip connections of the U-Net. Residual blocks could mitigate the gradient vanishing problem during training by adding the input of the module to its output, which also contributes to the speed of convergence [62]. In this configuration, the network can be built deeper. The work of [53, 59, 64–66] used residual connections or residual blocks instead of regular convolutions in their network architecture for robust segmentation of various brain structures. Dense blocks could strengthen feature propagation and encourage feature reuse to improve segmentation accuracy. However, they require more computational resources during training. Zhang et al. [46, 56] employed the Tiramisu network [67], a densely U-shaped network, to produce superior multiple sclerosis (MS) lesion segmentation.

The attention module is another commonly used tool in segmentation to focus on salient features [4]. It can be categorized into spatial attention and channel attention modules. Li et al. [53] use spatial attention modules in the skip connections for extracting smaller subcortical structures. Similarly, attention modules are used between skip connections and in the decoder part in the work of [51, 68] for segmenting vestibular schwannoma and cochlea. In addition, Zhang et al. [69] proposed to use slice-wise attention networks in 3D CNNs for MS segmentation. Applying the slice-

wise attention in three different orientations improves the computational efficiency compared to the regular attention module. Hou et al. [70] proposed the cross-attention block, which combines channel attention and spatial attention. Moreover, in [71], a skip attention unit is used for brain tumor segmentation. Zhou et al. [72] build fusion blocks based on the attention module. Attention modules have also been used for brain tumor segmentation [73].

Transformers As discussed in Subheading 2.1.2, transformers have become popular in medical image segmentation [74–76]. Transformers leverage the long-range dependencies and can better capture low-level details. In practice, they can replace CNNs [77], be combined with CNNs [78, 79], or integrated into CNNs [80]. Some recent works [14, 15, 77] have shown that the implementation of transformer on U-Net architecture can achieve superior performance in medical image segmentation compared to their CNN counterparts.

2.3.4 Framework Configuration

The single network mainly focuses on a single task during training and may ignore other potentially useful information. To improve the segmentation accuracy, frameworks with multiple encoders and decoders have been proposed [53, 81, 82].

Multi-task Networks As the name suggests, multi-task networks attempt to simultaneously tackle a main task as well as auxiliary tasks, rather than focusing on a single segmentation task. These networks usually contain a shared encoder and multiple decoders for multiple tasks, which could help deal with class imbalance (Fig. 9). Compared to a single-task network, the learning ability of the encoder is increased from same domain tasks (e.g., multiple tasks of multiple decoders), which could improve segmentation performance. Simultaneously learning multiple tasks could also improve model generalizability. McKinley et al. [81] leverage the information of additional tissue types to increase the accuracy of MS lesion segmentation. Another common multi-task setting is to introduce an auxiliary reconstruction task [57].

Cascaded Networks A cascaded network is a series of connected networks such that the input of each downstream network is the output from an upstream network (Fig. 10). For example, a coarse-to-fine segmentation strategy can be used to reduce the high computational cost of training for 3D images [50, 53]. In this scenario, an upstream network could take downsampled images as input to roughly locate the target structures, allowing the images to be cropped to the region of interest for the downstream network. The downstream network could then produce high-quality segmentation in full resolution. Another advantage of this approach

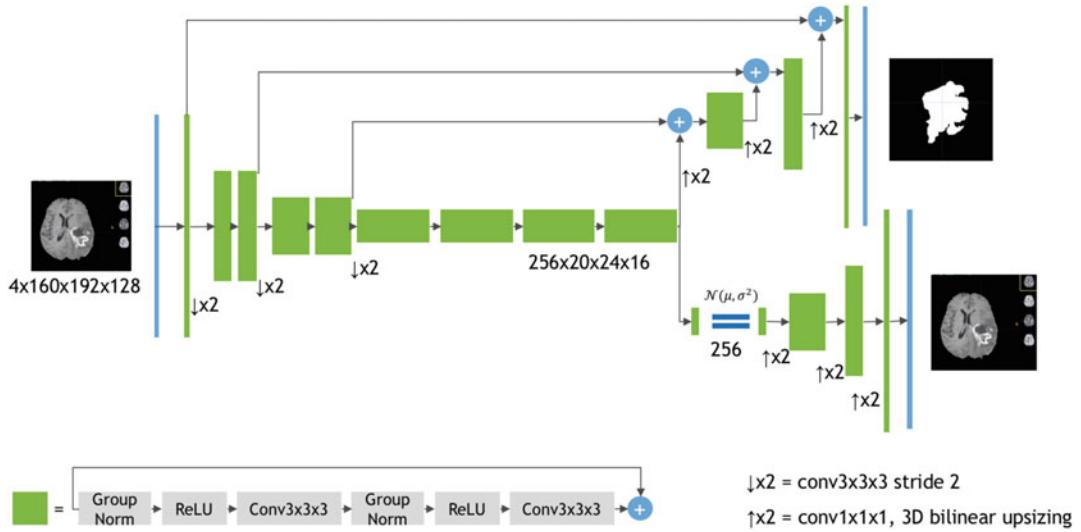


Fig. 9 Example of multi-task framework. The model takes four 3D MRI sequences (T1w, T1c, T2w, and FLAIR) as input. The U-Net structure (the top pathway with skip connection) serves as the segmentation network, and the output contains the segmentation maps of the three subregions (whole tumor (WT), tumor core (TC), and enhancing tumor (ET)). An auxiliary VAE branch (the bottom decoder) that reconstructs the input images is applied in the training stage to regularize the shared encoder. ©2019 Springer Nature. Reprinted, with permission, from [57]

is to reduce the impact of volume imbalance between foreground and background classes. However, the upstream network would determine the performance of the whole framework, and some global information is missing in the downstream networks.

Ensemble Networks To obtain a robust segmentation, a popular approach is to aggregate the output from multiple independent networks (i.e., no weights/parameters shared). Kanitsas et al. proposed the ensemble of multiple models and architectures (EMMA) [83] for brain tumor segmentation. Kao et al. [84] produce segmentation using 26 ensemble neural networks. Zhao et al. [85] proposed a framework for 3D segmentation with multiple 2D networks that take input from different views. Huo et al. [82] proposed the spatially localized atlas network tiles (SLANT) method to distribute multiple networks for 3D high-resolution whole-brain segmentation. Among their variants, SLANT-27 (Fig. 11), which ensembles 27 networks, produces the best result. Last but not least, many medical image segmentation challenge participants use model ensembling to achieve high performance.

2.3.5 Multiple Modalities and Timepoints

Many neuroimaging studies contain multiple modalities or multiple timepoints per subject. This additional information is clearly valuable and can be leveraged to improve segmentation performance.

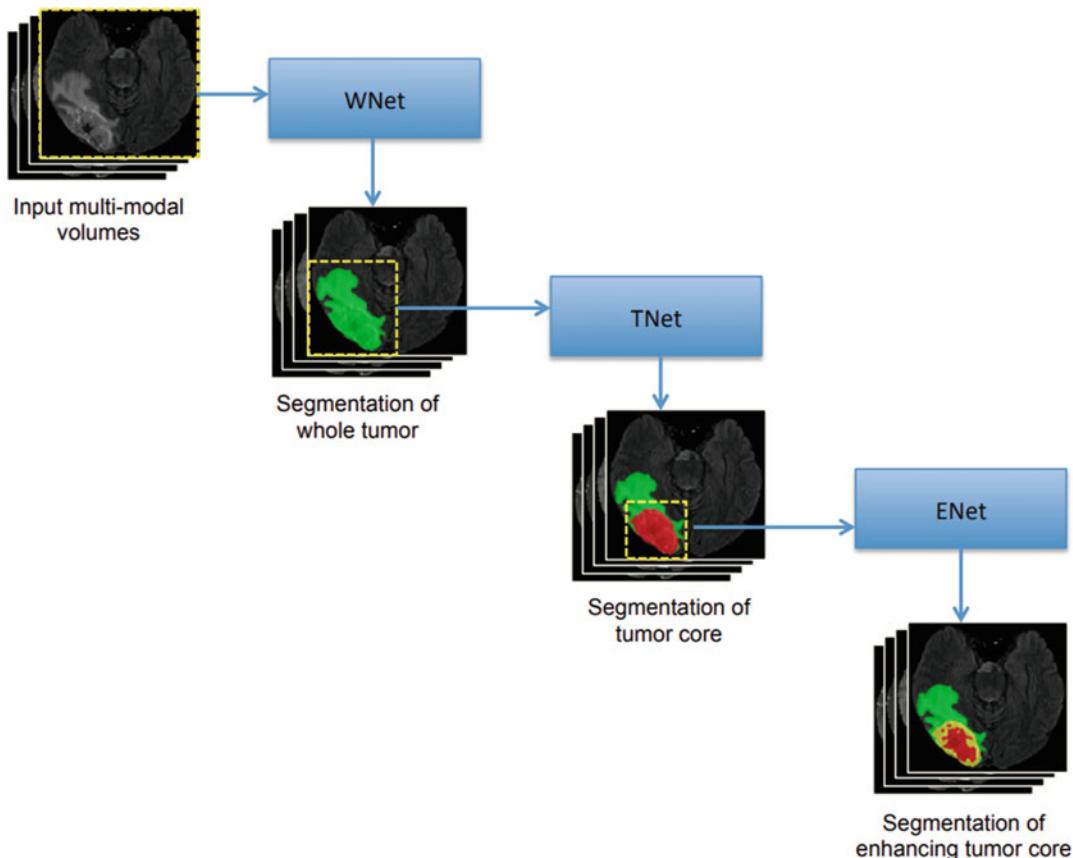


Fig. 10 Example of cascaded networks. WNet segments the whole tumor from the input multimodal 3D MRI. Then based upon the segmentation, a bounding box (yellow dash line) can be obtained and used to crop the input. The TNet takes the cropped image to segment the tumor core. Similarly, the ENet segments the enhancing tumor core by taking the cropped images determined by the segmentation from the previous stage. ©2018 Springer Nature. Reprinted, with permission, from [50]

Multiple Modalities Different imaging modalities offer different visualizations of various tissue types. Multi-modality datasets can be thus leveraged to improve segmentation accuracy. For example, Zhang et al. [86] proposed a framework with two independent networks that take two different modalities as inputs. Instead of combining single modality networks, Zhang et al. [46] concatenate multi-modality data as different channels of inputs. However, not all modalities are available in clinical practice: (1) the MRI sequences can vary between different imaging sites and (2) some modalities may be unusable due to poor image quality. This is known as the missing modality problem. To tackle this problem, Havaei et al. [87] proposed a deep learning method that is robust to missing modalities for brain tumor and MS segmentation, which contains an abstraction layer that transforms feature maps into statistics to help learning during training. In [88], the authors further improved modality dropout by introducing dynamic filters

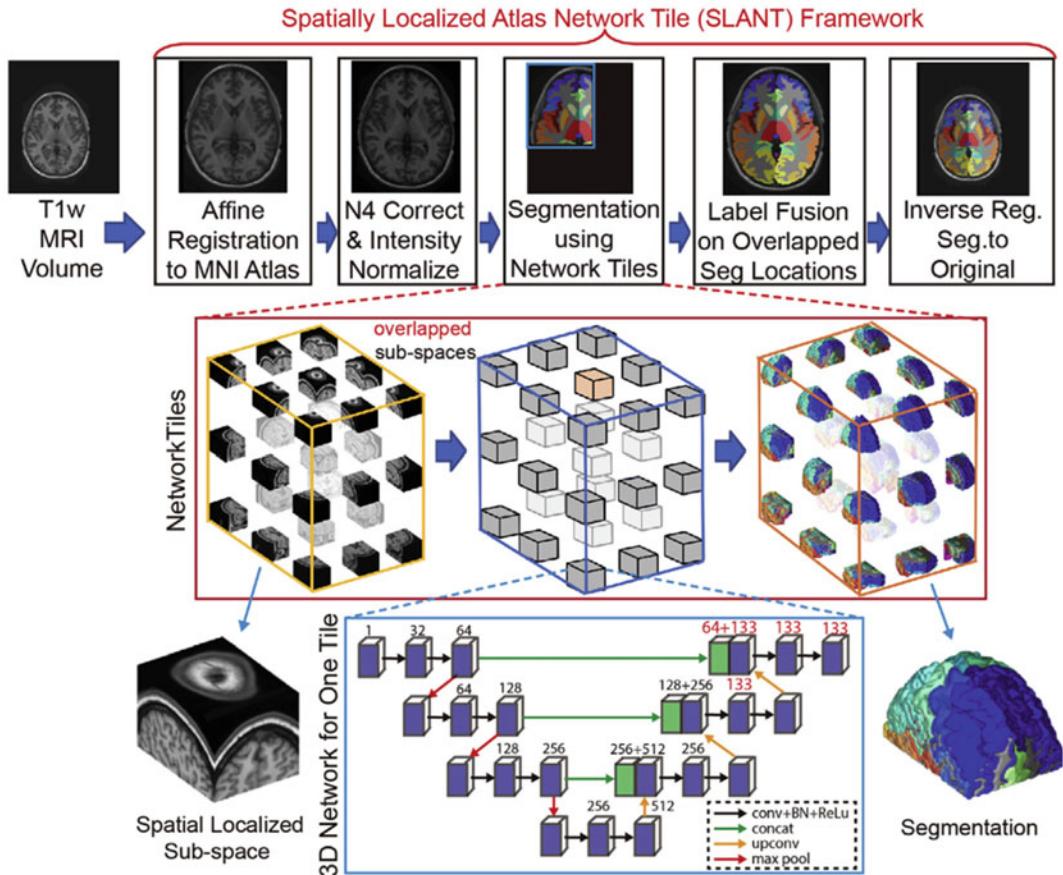


Fig. 11 SLANT-27: An example of ensemble networks. The whole brain is split into 27 overlapping subspaces with regard to their spatial locations (yellow cube). For each location, there is an independent 3D fully convolutional network (FCN) for segmentation (blue cube). The ensemble is achieved by label fusion on overlapping locations. ©2019 Elsevier. Reprinted, with permission, from [82]

and co-training strategy for MS lesion segmentation. In [89, 90], the authors used knowledge distillation scheme to transfer the knowledge from full-modality data to each missing condition with individual models.

Multiple Timepoints Data from multiple timepoints are important for tracking the longitudinal changes in a single subject. The additional timepoints can also be used as temporal context to improve the segmentation for each timepoint. In [45], longitudinal data are concatenated as a multichannel input to improve segmentation. In the work of [91], the stacked convolutional long short-term memory modules (C-LSTMs) are integrated into CNN for 4D medical image segmentation, which allows the model to learn the correlation and overall trends from longitudinal data. Li et al. [92] also proposed a framework with C-LSTM modules for segmenting longitudinal data jointly.

2.4 Semi-supervised Methods

2.4.1 Background

Given a considerable amount of labeled data, deep learning-based methods have achieved state-of-the-art performances in various medical image analysis applications. However, it is a laborious and time-consuming process to obtain dense pixel/voxel-level annotations for segmentation tasks. Since accurate annotations require expertise in medical domain, they are also expensive to collect. It is therefore desirable to leverage unlabeled data alongside the labeled data to improve model performance, an approach typically known as semi-supervised learning (SSL). Intuitively, these unlabeled data can provide critical information on the data distribution and thus can be used to improve model robustness by exploring this distribution.

Conceptually, SSL falls in between supervised learning (fully labeled data) and unsupervised learning (no labeled data). In SSL, we have access to both a labeled dataset $\mathcal{D}_L = \{(x_l^{(i)}, y_l^{(i)}) | i = 1, 2, \dots, n_L\}$, where $y_l^{(i)}$ is the i th manually annotated ground truth mask in the context of segmentation task, and an unlabeled dataset $\mathcal{D}_U = \{(x_u^{(i)} | i = 1, 2, \dots, n_u\}$. Typically, $n_u \gg n_L$. The main objective of SSL is to train a segmentation network X by leveraging both \mathcal{D}_L and \mathcal{D}_U to surpass the performances achieved by solely supervised learning with \mathcal{D}_L or unsupervised learning with \mathcal{D}_U .

According to [93], there are mainly three underlying assumptions held by SSL: (1) smoothness assumption, (2) low-density assumption, and (3) cluster assumption. The smoothness assumption states that the data points that are close by in the input or latent space should have similar or identical labels. With this assumption, we can expect the labels of unlabeled data to be similar to those of labeled data when these samples are similar in input or latent space, i.e., the labels from the labeled dataset can be transferred to the unlabeled dataset. In the low-density assumption, we assume that the decision boundary of a classifier should ideally not pass through the high density of the marginal data distribution. Placing the decision boundary in a high-density region would violate the smoothness assumption because the labels would be more likely to be dissimilar for similar data points. Lastly, the cluster assumption states that each cluster of data points should belong to the same class. This assumption is necessary because if the data points from the unlabeled and labeled datasets cannot be meaningfully clustered, the unlabeled data cannot be used to improve the model performance trained from only the labeled data.

2.4.2 Overview of Semi-supervised Techniques

In the semi-supervised learning literature, most of the techniques are originally designed and validated in the context of classification tasks. However, these methods can be readily adapted to segmentation tasks since a segmentation task can be viewed as pixel-wise classification. In this chapter, we mainly categorize the SSL approaches into three techniques, namely, (1) consistency

Table 1
Summary of classic semi-supervised learning methods

Method	Consistency regularization	Entropy minimization	Self-training
Pseudo-label [94]	No	Yes	Yes
Π model [95]	Yes	No	Yes
Temporal ensembling [95]	Yes	No	Yes
Mean teacher [96]	Yes	No	No
UDA [97]	Yes	Yes	No
MixMatch [98]	Yes	Yes	No
FixMatch [99]	Yes	Yes	No

regularization, (2) entropy minimization, and (3) self-training. However, most existing SSL approaches often employ a combination of these techniques rather than a single one, as summarized in Table 1. In the following sections, we will discuss each approach in detail and introduce some of the most important SSL techniques alongside.

2.4.3 Consistency Regularization

In semi-supervised learning, consistency regularization has been widely used as a technique to make use of unlabeled data. The idea of consistency regularization is based on the smoothness assumption that the network outputs should remain the same even if the input data is perturbed slightly (i.e., do not vary dramatically in the input space). The consistency between the predictions of an unlabeled sample and its perturbed counterpart can be used as a supervision mechanism for training to leverage the unlabeled data. In such scenarios, we can formulate the semi-supervised training objective as follows:

$$\ell_{SSL} = \sum_{x_l, y_l \in D_L} L_S(x_l, y_l) + \alpha \sum_{x_u \in D_U} L_C(x_u, \tilde{x}_u) \quad (31)$$

where L_S is the supervised loss for labeled data. For segmentation tasks, L_S can be one of the segmentation losses we presented in Subheading 2.1.3. x_u and \tilde{x}_u are the unlabeled data and its perturbed version, respectively. L_C is the consistency loss function. Mean squared error loss and KL divergence loss have been widely used as L_C in the SSL literature. α is a balancing term to weigh the impact of consistency loss from unlabeled data.

It is worth noting that the random permutations involved in consistency regularization can be implemented in different ways. For instance, the Π model [95] encourages consistent network outputs between two versions of the same input data, i.e., with different data augmentation and different network dropout

conditions. In this way, training can leverage the labeled data by optimizing the supervised segmentation loss and the unlabeled data by using this unsupervised consistency loss. In mean teacher [96], the authors propose to compute the consistency between the outputs of the student network and the teacher network (which uses the exponential moving average of the student network weights) from the same input data. In unsupervised data augmentation (UDA) [97], unlabeled data are augmented via different augmentation strategies such as RandAugment [100] and are fed to the same network to obtain two model predictions, which are used to compute the consistency loss. Similarly, in MixMatch [98], another very popular SSL method, an unlabeled image is augmented K times and the average of their outputs is sharpened, which is then used as the supervision signal to compute the consistency loss. Moreover, in FixMatch [99], the consistency loss is computed on the weakly and strongly augmented versions of the same input. In summary, consistency regularization has been widely used in various SSL techniques to leverage the unlabeled data.

Application: MTANS MTANS [101] is an SSL framework for brain lesion segmentation. As shown in Fig. 12, the MTANS framework is built upon the mean teacher model [96] where both the teacher and the student models are used to segment the brain lesions as well as the signed distance maps of the object surfaces. As a variant of the mean teacher model, MTANS incorporates **consistency regularization** in the training strategy. Specifically, the authors propose to compute the multi-scale feature consistency as consistency regularization, while the traditional mean teacher model only computes the consistency at the output level. Besides, a discriminator network is used to extract hierarchical features and differentiate the signed distance maps obtained by labeled and unlabeled data. In experiments, MTANS is evaluated on three public brain lesion datasets including ISBI 2015 (multiple sclerosis) [102], ISLES 2015 (ischemic stroke) [103], and BRATS 2018 (brain tumor) [104]. Experimental results show that MTANS can outperform the supervised baseline and other competing SSL methods when trained with the same amount of labeled data.

2.4.4 Entropy Minimization

Entropy minimization is another important SSL technique and is often used together with consistency training. Generally, entropy is the measure of the disorder or the uncertainty of a system. In the context of SSL, this term often refers to the uncertainty in the pseudo-label obtained by the unlabeled data. Entropy minimization, also known as minimum entropy regularization, aims to encourage the model to produce high-confidence predictions. The idea of entropy minimization is built upon the low-density assumption as it requires the network to output low-entropy

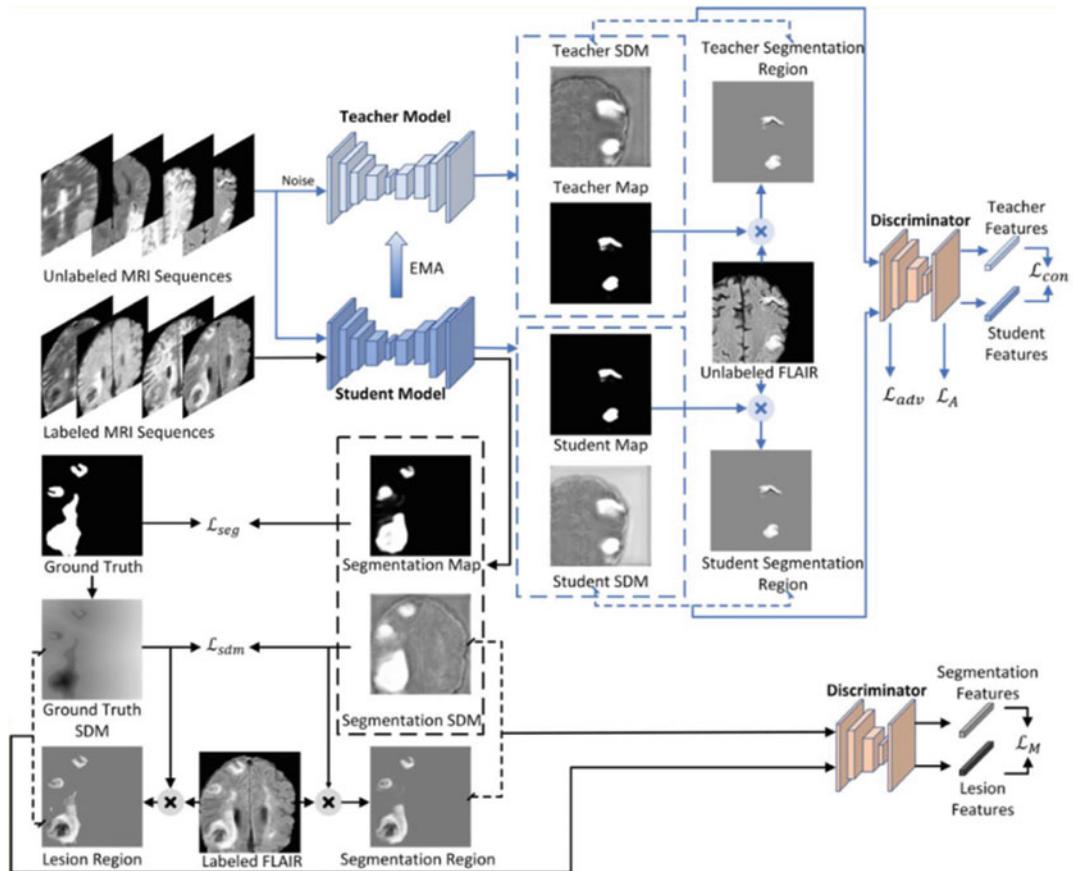


Fig. 12 An illustration of the MTANS framework. The blue solid lines indicate the path of unlabeled data, while the labeled data follows the black lines. The two segmentation models provide the segmentation map and the signed distance map (SDM). The discriminator is applied to check the consistency of the outputs from the teacher and student models. The parameters of the teacher model are updated according to the student model using the exponential moving average (EMA). ©2021 Elsevier. Reprinted, with permission, from [101]

predictions on unlabeled data. The high-confidence pseudo-labels have been found very effective when used as the supervision for unlabeled data. For example, in MixMatch, the pseudo-label of the unlabeled data, i.e., the average predictions of K augmented samples, is “sharpened” by adjusting the prediction distribution. This sharpening process is an implicit way to minimize the entropy on the unlabeled data distribution. In pseudo-label [94], the authors propose to construct the hard (one-hot) pseudo-labels from the high-confidence predictions of the unlabeled data, which is another form of entropy minimization. In addition, the UDA method proposes to compute the consistency loss only when the highest probability in the predicted class is above a pre-defined threshold. Similarly, in FixMatch, the predictions of the weakly augmented unlabeled data are first filtered by a pre-defined threshold and later converted to a one-hot pseudo-label.

2.4.5 Self-training

Self-training is an iterative training process where the network uses the high-confidence pseudo-labels of the unlabeled data from previous training steps. Interestingly, it has been shown that self-training is equivalent to a version of the classification EM algorithm [105]. The ideas of self-training and consistency regularization are very similar. Here, we differentiate these two concepts as follows: for consistency regularization, the supervision signals of the unlabeled data are generated online, i.e., from the current training epoch; in contrast, for self-training, the pseudo-labels of unlabeled data are generated offline, i.e., generated from the previous training epoch/epochs. Typically, in self-training, the pseudo-labels produced from previous epochs need to be carefully processed before being used as the supervision, as they are crucial to the effectiveness of the self-training methods. In the SSL literature, pseudo-label [94] is a representative method that uses self-training. In pseudo-label, the network is first trained on the labeled data only. Then the pseudo-labels of the unlabeled data are obtained by feeding them to the trained model. Next, the top K predictions on the unlabeled data are used as the pseudo-labels for the next epoch. The training objective function of pseudo-label is as follows:

$$L_{PL} = \sum_{x_l, y_l \in D_L} L_S(x_l, y_l) + \alpha(t) \sum_{x_u \in D_U} L_S(x_u, \tilde{y}_u) \quad (32)$$

where \tilde{y} is the pseudo-label and $\alpha(t)$ is a balancing term to weigh the importance of pseudo-label training. Particularly, $\alpha(t)$ is designed to slowly increase to help the optimization process to avoid poor local minima [94]. Note that both labeled and unlabeled data are trained in a supervised manner with ground truth labels y_l and pseudo labels \tilde{y}_u .

Application: 4S In this study, the authors propose a sequential semi-supervised segmentation (4S) framework [106] for serial electron microscopy image segmentation. As shown in Fig. 13, 4S relies on the **self-training** strategy as it applies pseudo-labeling to all slices in the target continuous images, with only a small number of consecutive input slices. Specifically, a few labeled samples are used for the first round of training. The trained model is then used to generate pseudo-labels for the next sample. Afterward, the segmentation model is retrained using the pseudo-labels and produces new pseudo-labels for the next slices. This method was evaluated on the ISBI 2012 dataset (neural cell membranes) [107] and Japanese carpenter ant dataset (nestmate discriminant sensory elements) [108]. Results show that 4S has achieved better performance than the supervised learning-based method.

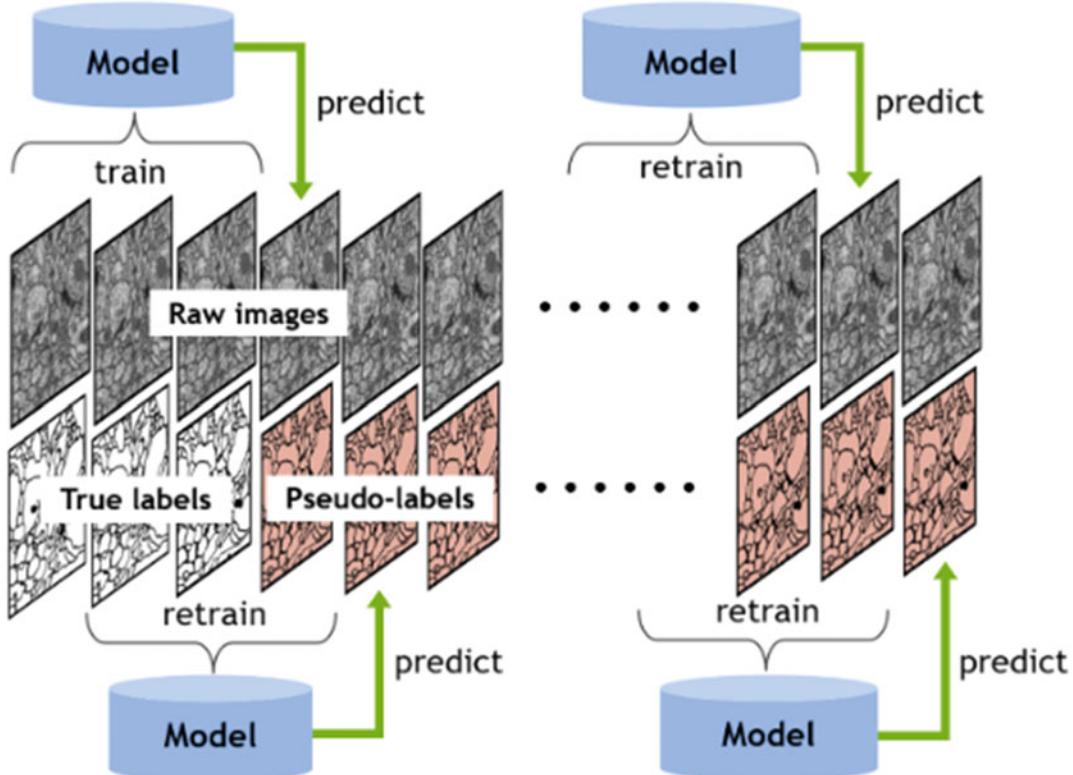


Fig. 13 The workflow of the 4S framework. Based on the assumption that consecutive images are strongly correlated, the manual annotations (true labels) are provided for the first few slices. These labeled data are used for the initial training. Then the model can provide the pseudo-labels for the next few slices which can be applied for retraining. Adapted from [106] (CC BY 4.0)

2.5 Unsupervised Methods

2.5.1 Background

As suggested in Subheadings 2.3 and 2.4, most deep segmentation models learn to map the input image x to the manually annotated ground truth y . Although semi-supervised approaches can drastically reduce the need for labels, low availability of ground truth is still a primary concern for the development of learning-based models. Another disadvantage of supervised learning approaches becomes evident when considering the anomaly detection/segmentation task: a model can only recognize anomalies that are similar to those in the training dataset and will likely fail with rare findings that may not appear in the training data [109].

Unsupervised anomaly detection (UAD) methods have been developed in recent years to tackle these problems. Since no ground truth labels are provided, the models are designed to capture the inherent discrepancy between healthy and pathological data distributions. *The general idea is to represent the distribution of normal brain anatomy by a deep model that is trained exclusively on healthy subjects* [109]. Consequently, the pathological subjects are out of

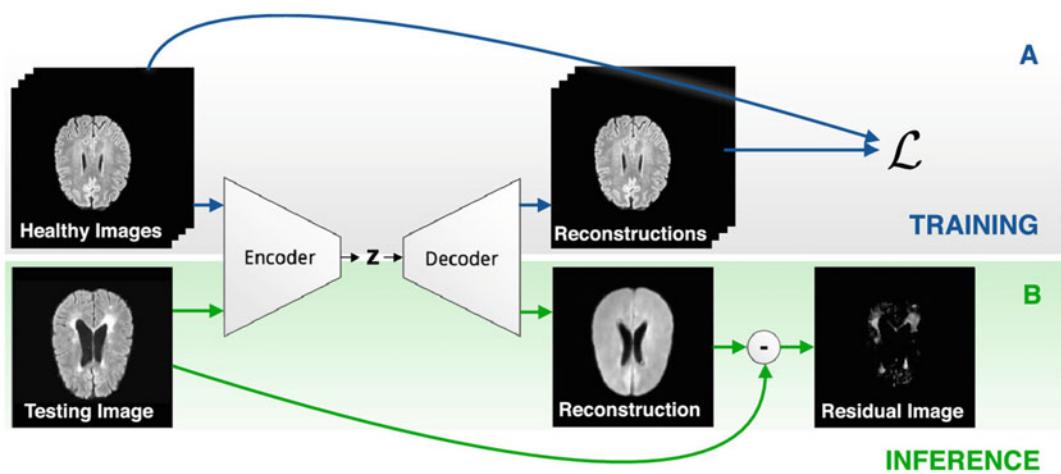


Fig. 14 The general idea of unsupervised anomaly detection (UAD) realized by an auto-encoder. (a) Train the model with only healthy subjects. (b) Test with pathological samples. The residual image depicts the anomalies. ©2021 Elsevier. Reprinted, with permission, from [109]

the distribution modeled by the network. Usually, this neural network has an encoder-decoder architecture such that the output will be a reconstruction of the input image. Since not well represented by the training data, the abnormal region cannot be fully reconstructed. Hence, the pixel-wise reconstruction error can be used as an estimate of the anomalous region. Figure 14 illustrates this process.

The auto-encoder (AE) and its variations (Fig. 15) are widely used in the UAD problem. All these models generate a low-dimensional representation of the input image termed latent vector z at the bottleneck. Most of the research concentrates on manipulating the distribution of z so that the abnormal region can be “cured” in the reconstruction. This process is often referred to as image restoration (or sometimes image inpainting) in the computer vision literature. The following sections will discuss some mainstream approaches categorized by the model structure implemented.

2.5.2 Auto-encoders

The auto-encoder (AE) (Fig. 15a) is the simplest encoder-decoder structure. Let an encoder f_θ and a decoder g_ϕ , where θ, ϕ are model parameters. Given a healthy input image $X^b \in \mathbb{R}^{D \times H \times W}$, the encoder learns to project it to a lower-dimensional latent space $z = f_\theta(X^b)$, $z \in \mathbb{R}^L$. Then the decoder recovers the original image from the latent vector as $\hat{X}^b = g_\phi(z)$. The model is trained by minimizing the loss function \mathcal{L} that delineates the difference between the input and the reconstructed image:

$$\underset{\theta, \phi}{\operatorname{argmin}} \mathcal{L}_{\theta, \phi}(X^b, \hat{X}^b) = \|X^b - \hat{X}^b\|_n \quad (33)$$

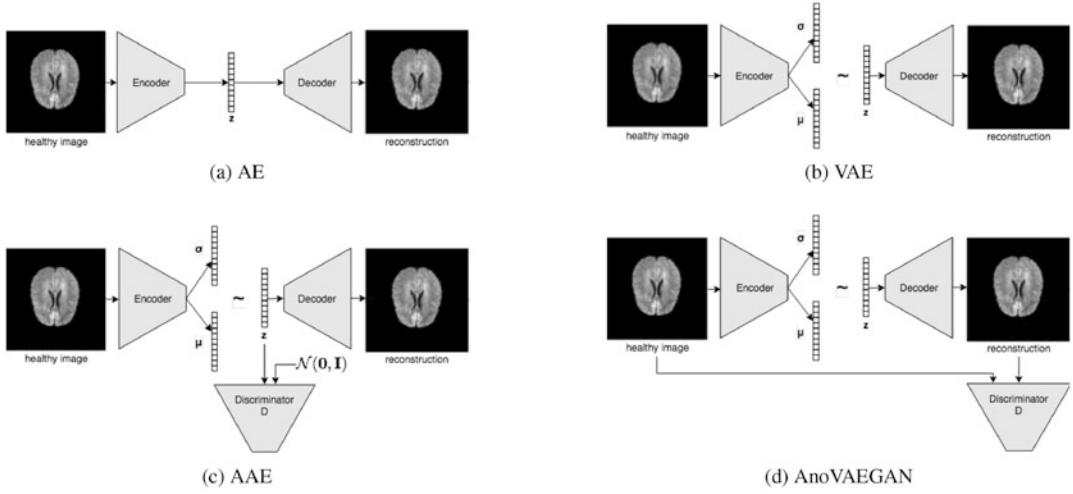


Fig. 15 Variations of auto-encoder. **(a)** The auto-encoder. **(b)** The variational auto-encoder. **(c)** The adversarial auto-encoder includes a discriminator that provides constraint on the distribution of the latent vector z . **(d)** Anomaly detection VAEGAN introduces a discriminator to check whether the reconstructed image lies in the same distribution as the healthy image. ©2021 Elsevier. Reprinted, with permission, from [109]

The ℓ_1 -norm ($n=1$) and ℓ_2 -norm (mean squared error) ($n=2$) are common choices for the loss function. The training stage is illustrated in Fig. 14a. When a sample with anomaly X^a is passed into the model, the abnormal region (e.g., lesion, tumor) cannot be well reconstructed in \hat{X}^a as the model has never seen the anomaly in the healthy training data. In other words, the AE-based methods leverage the models' dependence on training data to discern the region that is out of distribution. Figure 14b shows that the anomaly can be roughly represented by the reconstruction error $\hat{\mathbf{Y}} = |\mathbf{X}^a - \hat{\mathbf{X}}^a|$.

Bayesian Auto-encoder Pawlowski et al. [110] report a Bayesian convolutional auto-encoder to model the healthy data distribution. They introduce the model uncertainty and deem the reconstructed image as the Monte Carlo (MC) estimate. Let F_{Θ} be the auto-encoder model with weights Θ and \mathcal{D} the training dataset. Then, the MC estimation can be expressed as

$$F_{\Theta}(X) = \int P(X|\Theta)P(\Theta|\mathcal{D})d\Theta \approx \frac{1}{N} \sum_{i=1}^N F_{\Theta_i}(X) \quad (34)$$

where $\Theta_i \sim P(\Theta|\mathcal{D})$. In practice, the authors apply the MC-dropout to model the weight uncertainty. The segmentation is still obtained by setting a threshold on the reconstruction error, as in the vanilla auto-encoder.

2.5.3 Variational Auto-encoders

In some applications, instead of utilizing the lack of generalizability of the model, we want to modify the latent vector \mathbf{z} to further guarantee that the reconstructed testing image $\hat{\mathbf{X}}^a$ looks closer to a healthy subject. Then again, the residual between \mathbf{X}^a and $\hat{\mathbf{X}}^a$ is sufficient to highlight the anomalies in the image. Usually, such manipulation requires probabilistic modeling for the latent manifold. Hence, many applications use the variational auto-encoder (VAE) [111] as the backbone of the model (Fig. 15b).

As previously stated, we want the model to learn the distribution of healthy data $P(\mathbf{X}^b)$. In the encoder-decoder structure, we introduce a latent vector \mathbf{z} at the bottleneck which follows a given distribution $P(\mathbf{z})$. Usually, $P(\mathbf{z})$ is assumed to follow a normal distribution $\mathcal{N}(0, \mathbf{I})$. The encoder and decoder are expressed by the conditional probabilities $Q_\theta(\mathbf{z}|\mathbf{X}^b)$ and $P_\phi(\mathbf{X}^b|\mathbf{z})$, respectively. Then the target distribution is given by

$$P(\mathbf{X}^b) = \int P_\phi(\mathbf{X}^b|\mathbf{z})P(\mathbf{z})d\mathbf{z}. \quad (35)$$

In addition to the reconstruction loss (e.g., ℓ_1/ℓ_2 norm), the Kullback-Leibler (KL) divergence $D_{KL}[Q_\theta(\mathbf{z}|\mathbf{X}^b)\|P(\mathbf{z})]$ that measures the distance of two distributions is another objective function to minimize. This term provides a constraint on the latent manifold such that the feature vector \mathbf{z} can be stochastically sampled from a normal distribution. By modifying Eq. 13.35 and then applying Jensen's inequality, we get the evidence lower bound (ELBO) \mathcal{L} for the log-likelihood of the healthy data:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{\mathbf{z} \sim Q_\theta(\mathbf{z}|\mathbf{X}^b)}[\log P_\phi(\mathbf{X}^b|\mathbf{z})] - D_{KL}[Q_\theta(\mathbf{z}|\mathbf{X}^b)\|P(\mathbf{z})] \quad (36)$$

It has been proved that maximizing the $\log P(\mathbf{X}^b)$ is equivalent to maximizing its ELBO, so $-\mathcal{L}$ serves as an objective function to optimize parameters θ and ϕ in the VAE model. By leveraging the same idea in the AE-based methods, the neural networks f_θ and g_ϕ model the normal brain anatomy if the training data contains only the healthy subjects. The approaches using VAE take one more step to guarantee the abnormal region cannot be recovered in the output, that is, modify the latent vector \mathbf{z}^a of the anomalous input such that $\mathbf{z}^a \sim Q_\theta(\mathbf{z}|\mathbf{X}^b)$.

Given that healthy brains \mathbf{X}^b and subjects with anomaly \mathbf{X}^a are differently distributed, it is reasonable to assume that their latent manifolds $Q_\theta(\mathbf{z}|\mathbf{X}^b)$ and $Q_\theta(\mathbf{z}|\mathbf{X}^a)$ also vary. Suppose $\mathbf{z}^a = f_\theta(\mathbf{X}^a)$, then naturally, $\mathbf{z}^a \sim Q_\theta(\mathbf{z}|\mathbf{X}^a)$. If we can modify \mathbf{z}^a so that $\mathbf{z}^a \sim Q_\theta(\mathbf{z}|\mathbf{X}^b)$, then after passing through the decoder $P_\phi(\mathbf{X}^b|\mathbf{z})$, the reconstruction output of the model $\hat{\mathbf{X}}^a$ would belong in $P(\mathbf{X}^b)$. That is to say, the modification in the latent manifold "cures" the anomaly. It is then easy to identify the anomaly as the residual between the input and output. The core part of the process is how to "cure" the latent representation of abnormal input. Some common ways are reported in the following examples.

Distribution Constraint A straightforward way to force $\mathbf{z}^a \sim Q_\theta(\mathbf{z}|\mathbf{X}^b)$ is adding a specific loss function at the bottleneck. Chen et al. [112] propose an adversarial auto-encoder (AAE) shown in Fig. 15c. The encoder works as a generator that produces samples in the latent space, and an additional discriminator is trained to judge whether the sample is drawn from the normal distribution. It emphasizes that all the latent representations should follow $\mathcal{N}(0, \mathbf{I})$, whether the input is healthy or not.

Discrete Encoding Another solution is proposed by Pinaya et al. [113]. They implement the vector-quantized variational auto-encoder (VQ-VAE) [114] to obtain a discrete representation of the latent tensor $\mathbf{z} \in \mathbb{R}^{n_z \times b \times w}$. It can be regarded as a $b \times w$ image which contains a vector $\mathbf{v}_i \in \mathbb{R}^{n_z}$ at each image location, where $i = 1, 2, \dots, b \times w$. The quantization of \mathbf{z} is realized by a pretrained embedding space ($\mathbf{e}_j \in \mathbb{R}^{n_z}$, where $j = 1, 2, \dots, K$). It serves as a codebook from which we can always find a code \mathbf{e}_j that is closest to the given \mathbf{v}_i . Then by simply replacing the vector \mathbf{v}_i with the index of its closest counterpart in the codebook, a quantized latent image $\mathbf{z}_q \in \mathbb{R}^{b \times w}$ is obtained. Theoretically, the abnormal region is “cured” by using \mathbf{e}_j to approximate \mathbf{v}_i as the embedding space follows a fixed distribution. As usual, the residual between input and the reconstructed image $|\mathbf{X} - \hat{\mathbf{X}}|$ is used to find the anomaly.

Different Normative Prior Different from the vanilla VAE described above, Dilokthanakul et al. [115] propose a Gaussian mixture VAE (GMVAE) that replaces the unit multivariate Gaussian prior in the latent space with a Gaussian mixture model. GMVAE was used for brain UAD by You et al. [116]. Following the same idea of ruling out the anomaly in the latent space, they restore the image with anomaly using maximum a posteriori estimation given the Gaussian mixture model.

2.5.4 Variational Auto-encoders with Generative Adversarial Networks

A generative adversarial network (GAN) consists of two modules, a generator G and a discriminator D . Similar with the decoder in VAE, the generator G models the mapping from a latent vector to the image space $\mathbf{z} \mapsto \mathcal{X}$ where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. The discriminator D can be deemed as a trainable loss function that judges whether the generated image $G(\mathbf{z})$ is in the image space \mathcal{X} . Combining the GAN discriminator and the VAE backbone has become a common idea in UAD problems. More details on GANs can be found in Chap. 5.

We note that D can be used as an additional loss in either latent or image space. In the adversarial auto-encoder (AAE) discussed above, the discriminator works to check whether the latent vector is drawn from the multivariate normal distribution. In contrast, Buar et al. [117] propose the AnoVAEGAN (Fig. 15d) model, in which the discriminator is applied in the image space to check whether the reconstructed image lies in the distribution of healthy data.

3 Medical Image Segmentation Challenges

Medical image segmentation is affected by different aspects of the specific task, such as image quality, visibility of tissue boundaries, and the variability of the target structures. Moreover, each organ, anatomical structure, or lesion type has its own specificities, and a given method may perform well for a given target and worse for another. Therefore, many public challenges are held that target specific problems in an attempt to create benchmarks and attract new researchers into an application field.

In this section, we briefly introduce some of the popular medical image segmentation challenges related to neuroimages. Then, we focus on brain tumor and multiple sclerosis (MS) segmentation challenges and summarize the most competitive methods for each challenge to highlight examples of the concepts discussed in this chapter.

3.1 Popular Segmentation Challenges

Medical image segmentation challenges aim to find better solutions to certain tasks, and it also provides researchers with benchmark or baseline methods for future development. Furthermore, the developments are driven by the need to clinical problems.

Medical Segmentation Decathlon There are ten different segmentation tasks in the medical segmentation decathlon (MSD), and each task focuses on certain organ/structure [118]. Specifically, liver tumors, brain tumors, hippocampus, lung tumors, prostate, cardiac, pancreas tumors, colon cancer, hepatic vessels, and spleen are the focused organ of each task. Each task usually involves a different modality. For example, multimodal multisite MRI data are used for brain tumors, while liver tumors are studied from portal venous-phase CT data. The Dice score (DSC) and normalized surface distance are used as evaluation metrics due their well-known behavior. Instead of finding the state-of-the-art performance for each task, MSD aims to find generalizable methods.

crossMoDA These years, domain adaptation techniques are a hot topic in medical image segmentation field, and a new challenge for unsupervised cross-modality domain adaptation is held for researchers which is named as cross-modality domain adaptation (crossMoDA) for medical image segmentation [119]. Furthermore, it is the first large and multi-class benchmark for unsupervised domain adaptation to segment vestibular schwannoma (VS) and cochleas. In a short summary, crossMoDA consists of labeled and unlabeled datasets of T1-weighted and T2-weighted MRIs (T1-w and T2-w images are unpaired). It aims to segment the corresponding regions of interest in unlabeled T2-weighted MRIs by leveraging the information from unpaired and labeled T1-weighted MRIs.

3.2 Brain Tumor Segmentation Challenge

Brain tumor segmentation (BraTS) challenge is an annual challenge held since 2012 [104, 120–123]. The participants are provided with a comprehensive dataset that includes annotated, multisite, and multi-parametric MR images. It is worth noting that the dataset has increased from 30 cases to 2000 between 2012 and 2021 [123].

Brain tumor segmentation is a difficult task for a variety of reasons [124], including morphological and location uncertainty of tumor, class imbalance between foreground and background, and low contrast of MR images and annotation bias. BraTS focuses on segmentations for the enhancing tumor (ET), tumor core (TC), and whole tumor (WT). The Dice score, 95% Hausdorff distance, sensitivity, and specificity are used as evaluation metrics.

BraTS 2021 There are two tasks in BraTS 2021 and one of them is segmentation of brain tumor subregions (task 1) [123].

Dataset The BraTS 2021 competition comprises 8000 multi-parametric MR images from 2000 patients. The data split is 1251 cases for training, 219 cases for the validation phase, and 530 cases for final ranking, and ground truth labels are only provided to participants for the training set. The validation phase aims to help the participants examine their algorithm, and the results are shown on the public leaderboard. The dataset contains four MRI modalities per subject (Fig. 16): T1-w, post-contrast T1-w (T1Gd), T2-w, and T2-fluid-attenuated inversion recovery (T2-FLAIR).

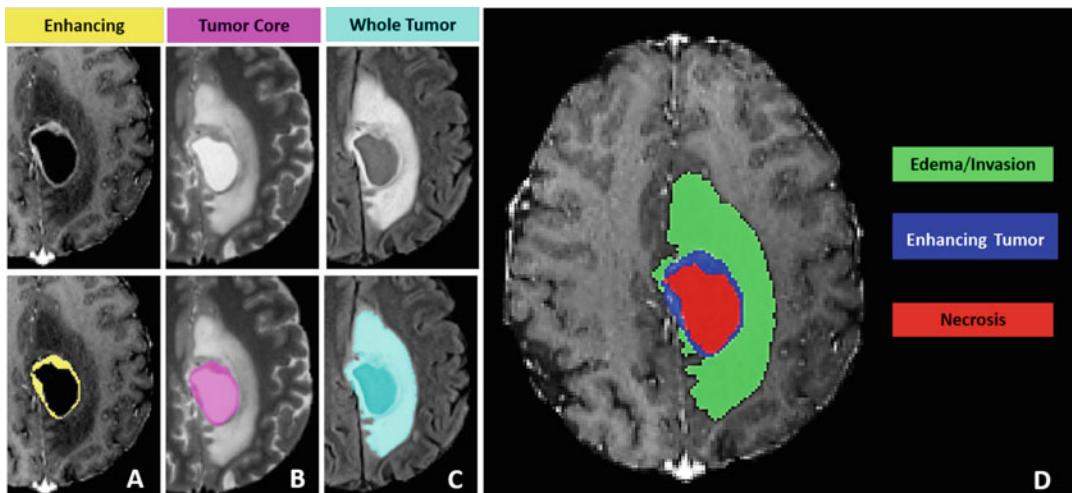


Fig. 16 BraTS 2021 dataset. The images and ground truth labels of enhancing tumor, tumor core, and whole tumor are shown in the panels A (T1w with gadolinium injection), B (T2w), and C (T2-FLAIR), respectively. Panel D shows the combined segmentations to generate the final tumor subregion labels. Replicated from [123] (CC BY 4.0)

The images were acquired at different institutions with different protocols and scanners. The pre-processing pipeline includes (1) co-registration to the same anatomical template, (2) resampling to isotropic 1mm^3 resolution, and (3) skull stripping.

Winner Method Luu et al. contributed a novel method [125] that won the first place in the final ranking after being applied to unseen test data. Their work is based on the nnU-Net, the winner of BraTS 2020. Some contributions include using group normalization instead of batch normalization; employing axial attention modules [126, 127] in the decoder part, which is efficient for multidimensional data; and building a deeper network. In the training phase, the networks were trained with 5-fold cross-validation. “Online” data augmentations were applied, including random rotation and scaling, elastic deformation, additive brightness augmentation, and gamma correction. The sum of the cross-entropy and Dice losses was used as the loss function. Last but not least, before feeding the input, the volumes were cropped to nonzero voxels and normalized by their mean and standard deviation.

3.3 Multiple Sclerosis Segmentation Challenge

Multiple sclerosis (MS) lesion segmentation from MR images is challenging for both radiologists and automated algorithms. The difficulties of this task include the large variability of lesion appearance, boundary, shape, and location, as well as variations in image appearance caused by different scanners and acquisition protocols from different institutes [128].

MSSEG-2 Delineation of new MS lesions on T2/FLAIR images is of interest as a biomarker of the effectiveness of anti-inflammatory disease-modifying drugs. Building upon the MSSEG (multiple sclerosis segmentation) challenge, MSSEG-2 (<https://portal.flair.iaria.fr/msseg-2/>) focuses on new MS lesion detection and segmentation. Here, we focus on the new lesion segmentation task.

Dataset The MSSEG-2 challenge dataset consists of 100 MS patients with 200 scans. Each subject has two FLAIR scans at different timepoints, with a time gap between 1 and 3 years. The images are acquired with 15 different 1.5T/3T scanners. Forty patients and their labels are used for training, and 120 scans of 60 patients are provided to test the performance.

Winner Method Zhang et al. proposed a novel method for segmentation of new MS lesions [56] that performed best for the Dice score evaluation. They adopted the model from [46], which is based on the U-Net and dense connections. The model inputs the concatenation of MR images from different timepoints and

outputs the new MS lesion segmentation for each patient. In addition, the 2.5D method, which stacks slices from three different orthogonal views (axial, sagittal, and coronal), is applied to each MR scan. In this way, both local and global information are provided to the model during training. Furthermore, to increase the generalizability of the model from the source domain to the target domain, three types of data augmentation are used that include image quality augmentation, image intensity augmentation, and spatial augmentation.

4 Conclusion

Image segmentation is a crucial task in medical image analysis. With the help of deep learning algorithms, one can achieve more precise segmentation on brain structures and lesions. In this chapter, we first introduced the fundamental components (Subheadings 2.1.1–2.1.6) needed to set up a complete deep neural network for a medical image segmentation task. Next, we provided a review of the rich literature on medical image segmentation methods categorized by supervision settings in Subheading 2.2–2.5. For each type of supervision, we explained the main ideas and provided example applications. Finally, we introduced some medical image segmentation challenges (Subheading 3) that have publicly available data, so that the readers can start their own projects.

References

1. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, Berlin, pp 234–241
2. Milletari F, Navab N, Ahmadi SA (2016) V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). IEEE, Piscataway, pp 565–571
3. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, et al (2018) Attention u-net: learning where to look for the pancreas. Preprint. arXiv:180403999
4. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D (2019) Attention gated networks: Learning to leverage salient regions in medical images. *Med Image Anal* 53:197–207
5. Isensee F, Jäger PF, Kohl SA, Petersen J, Maier-Hein KH (2019) Automated design of deep learning methods for biomedical image segmentation. Preprint. arXiv:190408128
6. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18(2):203–211
7. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, pp 424–432
8. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
9. Zhang J, Jiang Z, Dong J, Hou Y, Liu B (2020) Attention gate resU-Net for automatic MRI brain tumor segmentation. *IEEE Access* 8:58533–58545

10. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in Neural Information Processing Systems, pp 5998–6008
11. Luong MT, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. Preprint. arXiv:150804025
12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16×16 words: transformers for image recognition at scale. Preprint. arXiv:201011929
13. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Preprint. arXiv:210314030
14. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth HR, Xu D (2022) Unetr: transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 574–584
15. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y (2021) Transunet: transformers make strong encoders for medical image segmentation. Preprint. arXiv:210204306
16. Ma J, Chen J, Ng M, Huang R, Li Y, Li C, Yang X, Martel AL (2021) Loss odyssey in medical image segmentation. Med Image Anal 71:102035
17. Jadon S (2020) A survey of loss functions for semantic segmentation. In: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE, Piscataway, pp 1–7
18. Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M (2017) Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, Berlin, pp 240–248
19. Salehi SSM, Erdogmus D, Gholipour A (2017) Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging. Springer, Berlin, pp 379–387
20. El Jundi R, Petitjean C, Honeine P, Cheplygina V, Abdallah F (2021) High-level prior-based loss functions for medical image segmentation: a survey. Comput Vis Image Underst 210:103248
21. Maier-Hein L, Reinke A, Christodoulou E, Glocker B, Godau P, Isensee F, Kleesiek J, Kozubek M, Reyes M, Riegler MA, et al (2022) Metrics reloaded: pitfalls and recommendations for image analysis validation. Preprint. arXiv:220601653
22. Shattuck DW, Sandor-Leahy SR, Schaper KA, Rottenberg DA, Leahy RM (2001) Magnetic resonance image tissue classification using a partial volume model. NeuroImage 13(5): 856–876
23. Hahn HK, Peitgen HO (2000) The skull stripping problem in mri solved by a single 3D watershed transform. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, pp 134–143
24. Smith SM (2002) Fast robust automated brain extraction. Hum Brain Mapp 17(3): 143–155
25. Leung KK, Barnes J, Modat M, Ridgway GR, Bartlett JW, Fox NC, Ourselin S, Initiative ADN, et al (2011) Brain maps: an automated, accurate and robust brain extraction technique using a template library. NeuroImage 55(3):1091–1108
26. Ségonne F, Dale AM, Busa E, Glessner M, Salat D, Hahn HK, Fischl B (2004) A hybrid approach to the skull stripping problem in MRI. NeuroImage 22(3):1060–1075
27. Kleesiek J, Urban G, Hubert A, Schwarz D, Maier-Hein K, Bendszus M, Biller A (2016) Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. NeuroImage 129:460–469
28. Yogananda CGB, Wagner BC, Murugesan GK, Madhuranthakam A, Maldjian JA (2019) A deep learning pipeline for automatic skull stripping and brain segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, Piscataway, pp 727–731
29. Zhang Q, Wang L, Zong X, Lin W, Li G, Shen D (2019) Frnet: Flattened residual network for infant MRI skull stripping. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, Piscataway, pp 999–1002
30. Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, Wick A, Schlemmer HP, Heiland S, Wick W, et al (2019) Automated brain extraction of multisequence MRI using artificial neural networks. Hum Brain Mapp 40(17):4952–4964
31. Gao Y, Li J, Xu H, Wang M, Liu C, Cheng Y, Li M, Yang J, Li X (2019) A multi-view pyramid network for skull stripping on neonatal

- T1-weighted MRI. *Magn Reson Imaging* 63: 70–79
32. Li H, Zhu Q, Hu D, Gunnala MR, Johnson H, Sherbini O, Gavazzi F, D'Aiello R, Vanderver A, Long JD, et al (2022) Human brain extraction with deep learning. In: *Medical Imaging 2022: Image Processing*, vol 12032. SPIE, Bellingham, pp 369–375
 33. Kalavathi P, Prasath VS (2016) Methods on skull stripping of MRI head scan images—a review. *J Digit Imaging* 29(3):365–379
 34. Juntu J, Sijbers J, Van Dyck D, Gielen J (2005) Bias field correction for MRI images. In: *Computer Recognition Systems*. Springer, Berlin, pp 543–551
 35. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC (2010) N4itk: improved N3 bias correction. *IEEE Trans Med Imaging* 29(6):1310–1320
 36. Fortin JP, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, Roalf DR, Satterthwaite TD, Gur RC, Gur RE, et al (2017) Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* 161:149–170
 37. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp 2223–2232
 38. Reinholt JC, Dewey BE, Carass A, Prince JL (2019) Evaluating the impact of intensity normalization on MR image synthesis. In: *Medical Imaging 2019: Image Processing*, vol 10949. SPIE, Bellingham, pp 890–898
 39. Shinohara RT, Sweeney EM, Goldsmith J, Shiee N, Mateen FJ, Calabresi PA, Jarso S, Pham DL, Reich DS, Crainiceanu CM, et al (2014) Statistical normalization techniques for magnetic resonance imaging. *NeuroImage Clin* 6:9–19
 40. Brett M, Johnsrude IS, Owen AM (2002) The problem of functional localization in the human brain. *Nat Rev Neurosci* 3(3): 243–249
 41. Shin H, Kim H, Kim S, Jun Y, Eo T, Hwang D (2022) COSMOS: cross-modality unsupervised domain adaptation for 3D medical image segmentation based on target-aware domain translation and iterative self-training. Preprint. arXiv:220316557
 42. Dong H, Yu F, Zhao J, Dong B, Zhang L (2021) Unsupervised domain adaptation in semantic segmentation based on pixel alignment and self-training. Preprint. arXiv:210914219
 43. Liu H, Fan Y, Cui C, Su D, McNeil A, Dawant BM (2022) Unsupervised domain adaptation for vestibular schwannoma and cochlea segmentation via semi-supervised learning and label fusion. Preprint. arXiv:220110647
 44. Isensee F, Petersen J, Kohl SA, Jäger PF, Maier-Hein KH (2019) nnU-Net: breaking the spell on successful medical image segmentation. Preprint 1:1–8. arXiv:190408128
 45. Birenbaum A, Greenspan H (2016) Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks. In: *Deep Learning and Data Labeling for Medical Applications*. Springer, Berlin, pp 58–67
 46. Zhang H, Valcarcel AM, Bakshi R, Chu R, Bagnato F, Shinohara RT, Hett K, Oguz I (2019) Multiple sclerosis lesion segmentation with tiramisu and 2.5 D stacked slices. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Berlin, pp 338–346
 47. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B (2017) Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal* 36:61–78
 48. Dolz J, Desrosiers C, Ayed IB (2018) 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage* 170:456–470
 49. Li H, Zhang H, Hu D, Johnson H, Long JD, Paulsen JS, Oguz I (2020) Generalizing MRI subcortical segmentation to neurodegeneration. In: *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*. Springer, Berlin, pp 139–147
 50. Wang G, Li W, Ourselin S, Vercauteren T (2017) Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In: *International MICCAI Brainlesion Workshop*. Springer, Berlin, pp 178–190
 51. Li H, Hu D, Zhu Q, Larson KE, Zhang H, Oguz I (2021) Unsupervised cross-modality domain adaptation for segmenting vestibular schwannoma and cochlea with data augmentation and model ensemble. Preprint. arXiv:210912169
 52. Zhang L, Wang X, Yang D, Sanford T, Harmon S, Turkbey B, Wood BJ, Roth H, Myronenko A, Xu D, et al (2020) Generalizing deep learning for medical image segmentation to unseen domains via deep

- stacked transformation. *IEEE Trans Med Imaging* 39(7):2531–2540
53. Li H, Zhang H, Johnson H, Long JD, Paulsen JS, Oguz I (2021) MRI subcortical segmentation in neurodegeneration with cascaded 3D CNNs. In: Medical Imaging 2021: Image Processing, International Society for Optics and Photonics, vol 11596, p 115960W
 54. Dong H, Yang G, Liu F, Mo Y, Guo Y (2017) Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. In: Annual Conference on Medical Image Understanding and Analysis. Springer, Berlin, pp 506–517
 55. Beers A, Chang K, Brown J, Sartor E, Mammen C, Gerstner E, Rosen B, Kalpathy-Cramer J (2017) Sequential 3D u-nets for biologically-informed brain tumor segmentation. Preprint. arXiv:170902967
 56. Zhang H, Li H, Oguz I (2021) Segmentation of new MS lesions with tiramisu and 2.5 D stacked slices. In: MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure, p 61
 57. Myronenko A (2018) 3D MRI brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop. Springer, Berlin, pp 311–320
 58. Pérez-García F, Sparks R, Ourselin S (2021) Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput Methods Prog Biomed* 208:106236
 59. Kamnitsas K, Ferrante E, Parisot S, Ledig C, Nori AV, Criminisi A, Rueckert D, Glocker B (2016) Deepmedic for brain tumor segmentation. In: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer, Berlin, pp 138–149
 60. Haweai M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, Pal C, Jodoin PM, Larochelle H (2017) Brain tumor segmentation with deep neural networks. *Med Image Anal* 35:18–31
 61. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3431–3440
 62. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
 63. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4700–4708
 64. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH (2017) Brain tumor segmentation and radiomics survival prediction: contribution to the brats 2017 challenge. In: International MICCAI Brainlesion Workshop. Springer, Berlin, pp 287–297
 65. Chang PD (2016) Fully convolutional deep residual neural networks for brain tumor segmentation. In: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer, Berlin, pp 108–118
 66. Castillo LS, Daza LA, Rivera LC, Arbeláez P (2017) Volumetric multimodality neural network for brain tumor segmentation. In: 13th International Conference on Medical Information Processing and Analysis, vol 10572. International Society for Optics and Photonics, Bellingham, p 105720E
 67. Jégou S, Drozdal M, Vazquez D, Romero A, Bengio Y (2017) The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 11–19
 68. Wang G, Shapey J, Li W, Dorent R, Demitriadis A, Bisdas S, Paddick I, Bradford R, Zhang S, Ourselin S, et al (2019) Automatic segmentation of vestibular schwannoma from T2-weighted MRI by deep spatial attention with hardness-weighted loss. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, pp 264–272
 69. Zhang H, Zhang J, Zhang Q, Kim J, Zhang S, Gauthier SA, Spincemaille P, Nguyen TD, Sabuncu M, Wang Y (2019) RSANet: recurrent slice-wise attention network for multiple sclerosis lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, pp 411–419
 70. Hou B, Kang G, Xu X, Hu C (2019) Cross attention densely connected networks for multiple sclerosis lesion segmentation. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, Piscataway, pp 2356–2361
 71. Islam M, Vibashan V, Jose VJM, Wijethilake N, Utkarsh U, Ren H (2019) Brain tumor segmentation and survival prediction using 3D attention UNet. In:

- International MICCAI Brainlesion Workshop. Springer, Berlin, pp 262–272
72. Zhou T, Ruan S, Guo Y, Canu S (2020) A multi-modality fusion network based on attention mechanism for brain tumor segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, Piscataway, pp 377–380
 73. Sinha A, Dolz J (2020) Multi-scale self-guided attention for medical image segmentation. *IEEE J Biomed Health Inform* 25(1): 121–130
 74. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H (2022) Transformers in medical imaging: a survey. Preprint. arXiv:220109873
 75. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth H, Xu D (2022) Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. Preprint. arXiv:220101266
 76. Peiris H, Hayat M, Chen Z, Egan G, Harandi M (2021) A volumetric transformer for accurate 3D tumor segmentation. Preprint. arXiv:211113300
 77. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M (2021) Swin-UNET: Unet-like pure transformer for medical image segmentation. Preprint. arXiv:210505537
 78. Zhang Y, Liu H, Hu Q (2021) Transfuse: fusing transformers and CNNs for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, pp 14–24
 79. Li H, Hu D, Liu H, Wang J, Oguz I (2022) Cats: complementary CNN and transformer encoders for segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). IEEE, Piscataway, pp 1–5
 80. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM (2021) Medical transformer: gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, pp 36–46
 81. McKinley R, Wepfer R, Aschwanden F, Grunder L, Muri R, Rummel C, Verma R, Weisstanner C, Reyes M, Salmen A, et al (2019) Simultaneous lesion and neuroanatomy segmentation in multiple sclerosis using deep neural networks. Preprint. arXiv:190107419
 82. Huo Y, Xu Z, Xiong Y, Aboud K, Parvathaneni P, Bao S, Bermudez C, Resnick SM, Cutting LE, Landman BA (2019) 3D whole brain segmentation using spatially localized atlas network tiles. *NeuroImage* 194: 105–119
 83. Kamnitsas K, Bai W, Ferrante E, McDonagh S, Sinclair M, Pawlowski N, Rajchl M, Lee M, Kainz B, Rueckert D, et al (2017) Ensembles of multiple models and architectures for robust brain tumour segmentation. In: International MICCAI Brainlesion Workshop. Springer, Berlin, pp 450–462
 84. Kao PY, Ngo T, Zhang A, Chen JW, Manjuth B (2018) Brain tumor segmentation and tractographic feature extraction from structural MR images for overall survival prediction. In: International MICCAI Brainlesion Workshop. Springer, Berlin, pp 128–141
 85. Zhao X, Wu Y, Song G, Li Z, Zhang Y, Fan Y (2017) 3D brain tumor segmentation through integrating multiple 2D FCNNs. In: International MICCAI Brainlesion Workshop. Springer, Berlin, pp 191–203
 86. Zhang D, Huang G, Zhang Q, Han J, Han J, Yu Y (2021) Cross-modality deep feature learning for brain tumor segmentation. *Pattern Recogn* 110:107562
 87. Havaei M, Guizard N, Chapados N, Bengio Y (2016) Hemis: hetero-modal image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, pp 469–477
 88. Liu H, Fan Y, Li H, Wang J, Hu D, Cui C, Lee HZ Ho Hin, Oguz I (2022) Moddrop++: a dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities. Preprint. arXiv:220304959
 89. Wang Y, Zhang Y, Liu Y, Lin Z, Tian J, Zhong C, Shi Z, Fan J, He Z (2021) ACN: adversarial co-training network for brain tumor segmentation with missing modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, pp 410–420
 90. Azad R, Khosravi N, Merhof D (2022) SMU-Net: style matching U-Net for brain tumor segmentation with missing modalities. Preprint. arXiv:220402961
 91. Gao Y, Phillips JM, Zheng Y, Min R, Fletcher PT, Gerig G (2018) Fully convolutional structured LSTM networks for joint 4D medical image segmentation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, Piscataway, pp 1104–1108

92. Li H, Zhang H, Johnson H, Long JD, Paulsen JS, Oguz I (2021) Longitudinal subcortical segmentation with deep learning. In: Medical Imaging 2021: Image Processing, International Society for Optics and Photonics, vol 11596, p 115960D
93. Van Engelen JE, Hoos HH (2020) A survey on semi-supervised learning. *Mach Learn* 109(2):373–440
94. Lee DH, et al (2013) Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML, vol 3, p 896
95. Laine S, Aila T (2016) Temporal ensembling for semi-supervised learning. Preprint. arXiv:161002242
96. Tarvainen A, Valpola H (2017) Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. Preprint. arXiv:170301780
97. Xie Q, Dai Z, Hovy E, Luong MT, Le QV (2019) Unsupervised data augmentation for consistency training. Preprint. arXiv:190412848
98. Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel C (2019) Mixmatch: a holistic approach to semi-supervised learning. Preprint. arXiv:190502249
99. Sohn K, Berthelot D, Li CL, Zhang Z, Carlini N, Cubuk ED, Kurakin A, Zhang H, Raffel C (2020) Fixmatch: simplifying semi-supervised learning with consistency and confidence. Preprint. arXiv:200107685
100. Cubuk ED, Zoph B, Shlens J, Le QV (2020) Randaugment: practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 702–703
101. Chen G, Ru J, Zhou Y, Rekik I, Pan Z, Liu X, Lin Y, Lu B, Shi J (2021) MTANS: multi-scale mean teacher combined adversarial network with shape-aware embedding for semi-supervised brain lesion segmentation. *NeuroImage* 244:118568
102. Carass A, Roy S, Jog A, Cuzzocreo JL, McGrath E, Gherman A, Button J, Nguyen J, Prados F, Sudre CH, et al (2017) Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage* 148:77–102
103. Maier O, Menze BH, von der Gabeltz J, Häni L, Heinrich MP, Liebrand M, Winzeck S, Basit A, Bentley P, Chen L, et al (2017) ISLES 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Med Image Anal* 35:250–269
104. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, et al (2014) The multimodal brain tumor image segmentation benchmark (BraTS). *IEEE Trans Med Imaging* 34(10):1993–2024
105. Amini MR, Gallinari P (2002) Semi-supervised logistic regression. In: ECAI, vol 2, p 11
106. Takaya E, Takeichi Y, Ozaki M, Kurihara S (2021) Sequential semi-supervised segmentation for serial electron microscopy image with small number of labels. *J Neurosci Methods* 351:109066
107. Arganda-Carreras I, Turaga SC, Berger DR, Cireşan D, Giusti A, Gambardella LM, Schmidhuber J, Laptev D, Dwivedi S, Buhmann JM, et al (2015) Crowdsourcing the creation of image segmentation algorithms for connectomics. *Front Neuroanat* 9:142
108. Takeichi Y, Uebi T, Miyazaki N, Murata K, Yasuyama K, Inoue K, Suzuki T, Kubo H, Kajimura N, Takano J, et al (2018) Putative neural network within an olfactory sensory unit for nestmate and non-nestmate discrimination in the Japanese carpenter ant: the ultra-structures and mathematical simulation. *Front Cell Neurosci* 12:310
109. Baur C, Denner S, Wiestler B, Navab N, Albarqouni S (2021) Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. *Med Image Anal*, p 101952
110. Pawłowski N, Lee MC, Rajchl M, McDonagh S, Ferrante E, Kamnitsas K, Cooke S, Stevenson S, Khetani A, Newman T, et al (2018) Unsupervised lesion detection in brain CT using bayesian convolutional autoencoders. *MIDL*
111. Kingma DP, Welling M (2013) Auto-encoding variational bayes. Preprint. arXiv:13126114
112. Chen X, Konukoglu E (2018) Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. Preprint. arXiv:180604972
113. Pinaya WHL, Tudosi PD, Gray R, Rees G, Nachev P, Ourselin S, Cardoso MJ (2021) Unsupervised brain anomaly detection and segmentation with transformers. Preprint. arXiv:210211650
114. Van Den Oord A, Vinyals O, et al (2017) Neural discrete representation learning. *Adv Neural Inf Proces Syst* 30

115. Dilokthanakul N, Mediano PA, Garnelo M, Lee MC, Salimbeni H, Arulkumaran K, Shanahan M (2016) Deep unsupervised clustering with gaussian mixture variational autoencoders. Preprint. arXiv:161102648
116. You S, Tezcan KC, Chen X, Konukoglu E (2019) Unsupervised lesion detection via image restoration with a normative prior. In: International Conference on Medical Imaging with Deep Learning, PMLR, pp 540–556
117. Baur C, Wiestler B, Albarqouni S, Navab N (2018) Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In: International MICCAI Brainlesion Workshop. Springer, Berlin, pp 161–169
118. Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, Litjens G, Menze B, Ronneberger O, Summers RM, et al (2022) The medical segmentation decathlon. Nat Commun 13(1):1–13
119. Dorent R, Kujawa A, Ivory M, Bakas S, Rieke N, Joutard S, Glockner B, Cardoso J, Modat M, Batmanghelich K, et al (2022) Crossmoda 2021 challenge: benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. Preprint. arXiv:220102831
120. Ghaffari M, Sowmya A, Oliver R (2019) Automated brain tumor segmentation using multimodal brain scans: a survey based on models submitted to the BraTS 2012–2018 challenges. IEEE Rev Biomed Eng 13:156–168
121. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, Shinohara RT, Berger C, Ha SM, Rozycski M, et al (2018) Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BraTS challenge. Preprint. arXiv:181102629
122. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycski M, Kirby J, Freymann J, Farahani K, Davatzikos C (2017) Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. Cancer Imaging Arch 286
123. Baid U, Ghodasara S, Mohan S, Bilello M, Calabrese E, Colak E, Farahani K, Kalpathy-Cramer J, Kitamura FC, Pati S, et al (2021) The RSNA-ASNR-MICCAI braTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. Preprint. arXiv:210702314
124. Liu Z, Chen L, Tong L, Zhou F, Jiang Z, Zhang Q, Shan C, Wang Y, Zhang X, Li L, et al (2020) Deep learning based brain tumor segmentation: a survey. Preprint. arXiv:200709479
125. Luu HM, Park SH (2021) Extending nn-Unet for brain tumor segmentation. Preprint. arXiv:211204653
126. Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen LC (2020) Axial-deeplab: stand-alone axial-attention for panoptic segmentation. In: European Conference on Computer Vision. Springer, Berlin, pp 108–126
127. Ho J, Kalchbrenner N, Weissenborn D, Salimans T (2019) Axial attention in multidimensional transformers. Preprint. arXiv:191212180
128. Zhang H, Oguz I (2020) Multiple sclerosis lesion segmentation—a survey of supervised CNN-based methods. Preprint. arXiv:201208317

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 14

Image Registration: Fundamentals and Recent Advances Based on Deep Learning

Min Chen, Nicholas J. Tustison, Rohit Jena, and James C. Gee

Abstract

Registration is the process of establishing spatial correspondences between images. It allows for the alignment and transfer of key information across subjects and atlases. Registration is thus a central technique in many medical imaging applications. This chapter first introduces the fundamental concepts underlying image registration. It then presents recent developments based on machine learning, specifically deep learning, which have advanced the three core components of traditional image registration methods—the similarity functions, transformation models, and cost optimization. Finally, it describes the key application of these techniques to brain disorders.

Key words Image registration, Alignment, Atlas

1 Introduction

In medical image analysis, the *correspondence* between important features or analogous anatomy in two images is an important piece of information that can be used to study disease. Knowing the correspondences between spatial locations allows for comparisons between specific anatomical structures in the images. This allows us to answer questions such as “Is this structure larger in subject A than in subject B?” or “Is that structure malformed relative to the average population?” Likewise, knowing correspondences across time allows us to study changes in rates of disease processes. For example, “Is a disease causing the structure to grow or shrink over time?” or “How does the rate of change compare to an healthy individual?”

Correspondences between images also provide the ability to transfer information, which can be used as prior knowledge for tasks such as segmentation. Knowing the boundary for a specific anatomical structure in image A allows the image to be used as an atlas for finding those same boundaries in other images. If the correspondences between images A and B are known, then the

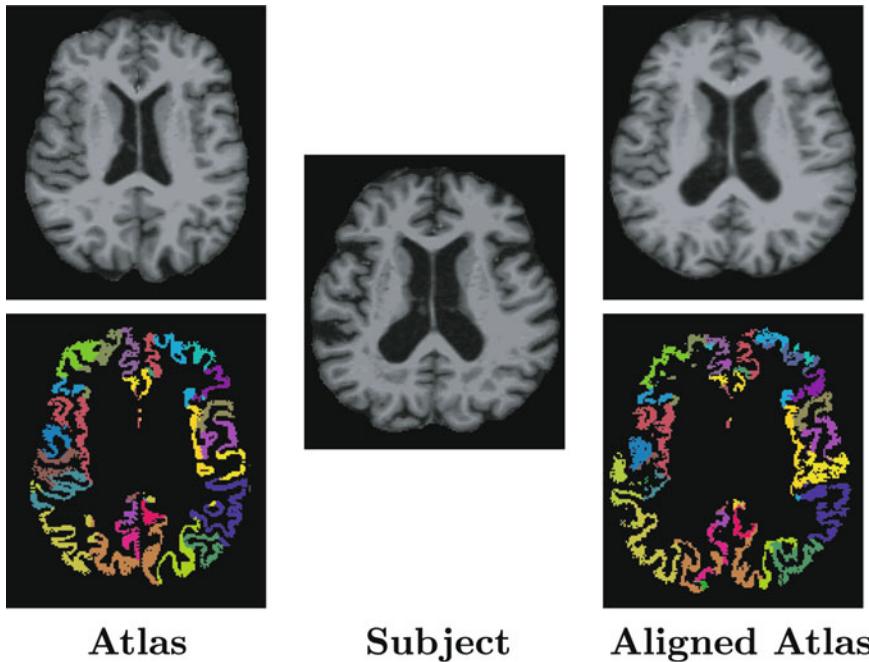


Fig. 1 Shown is an example of an atlas alignment using image registration between two different brain magnetic resonance images. The atlas image (top left) is transformed (top right) to be aligned with the fixed subject image (center). The transformation allows the anatomical labels from the atlas (bottom left) to be directly transferred (bottom right) to label the subject image

boundary in image A can be transferred through the correspondences and used as an approximate starting point for finding the analogous boundaries in image B (called the fixed image).

In the field of medical imaging and computer vision, the task of computing and aligning correspondences between different images is referred to as *image registration*. Given two images, image registration algorithms use image features such as image intensities or structures in the images to find a transformation that best aligns the correspondences between the two images. In Fig. 1, we show an example where such an algorithm is used to align the image intensities between two different brain images. We see that this alignment allows the anatomical labels on an atlas image to be directly transferred to the fixed image.

While the primary concept of image registration is simple, finding the solution is not so straightforward. The subject has been studied extensively for the past 40 years [1], and there is still little of consensus on the best general approach for the problem. We often cannot determine what are the correct correspondences between two images. In addition, we rarely know the exact way to model the transformation that best aligns those correspondences. We see from the example in Fig. 1 that aligning the intensity correspondences does not accurately align all of the anatomical correspondences between the images.

The number of varieties and applications of image registration that have been presented to date is tremendous [2, 3]. In this chapter, we will only discuss a limited subset of these techniques, specifically methods that have been developed in recent years that leverages machine learning (and in particular, deep CNNs) to solve the problem. We will start by providing a brief introduction to the fundamental building blocks of traditional image registration techniques and then delve into how various pieces of these designs have been developed and improved upon using machine learning models.

2 Fundamentals of Image Registration

The main goal of an image registration algorithm is to take a *moving image* and transform it to be spatially or temporally aligned with a target *fixed image*. The algorithm is generally defined by two parts: the type of transformation allowed to be performed on the moving image (the *transformation model*) and a definition of good alignment (the *similarity cost function*) between the two images. The algorithm is often iterative, in which case there is also an *optimizer*, which searches for how to adjust the transformation to best minimize the cost function. This is typically performed by estimating a transformation using the model, applying it to the moving image, and then evaluating the cost function between the transformed moving image and the fixed image. This cost then informs the algorithm on how to estimate a more accurate transformation for the next iteration. The process is repeated and optimized until either the moving and fixed images are considered aligned (i.e., a local minimum is reached in the cost function) or a maximum iteration count is exceeded. Figure 2 summarizes this iterative framework as a block diagram. Figure 3 shows several examples of registration results when using different transformation models to register between two MR images of the brain.

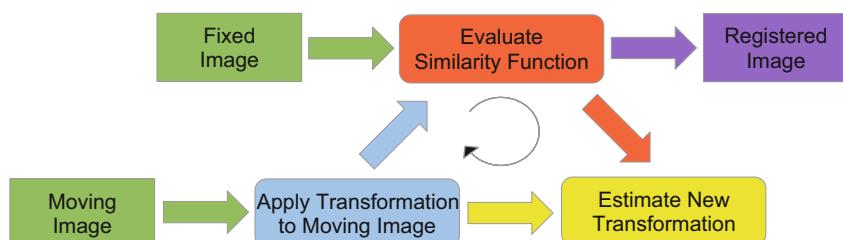


Fig. 2 Block diagram of the general registration framework. The coloring represents the main pieces of the framework: the input images (green), the output image (purple), the similarity cost function (orange), the transformation model (blue), and the optimizer (yellow)

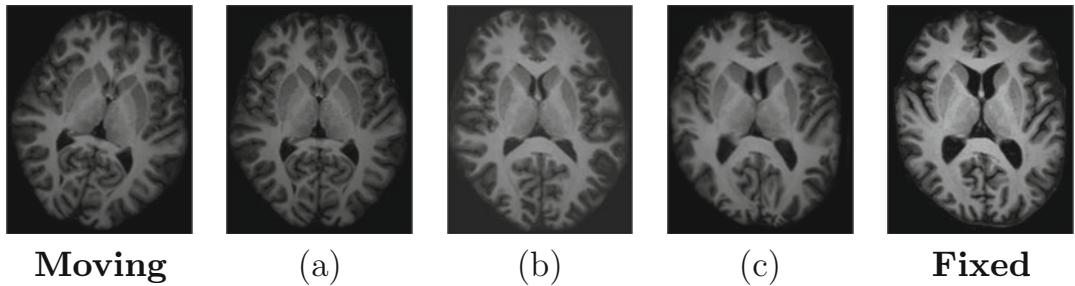


Fig. 3 Shown are examples of registration results between a moving and fixed MR image of the brain from two different subjects, using a **(a)** rigid, **(b)** affine, and **(c)** deformable registration

2.1 Registration as a Minimization Problem

To describe the general registration problem, we begin by using functions $S(\mathbf{x}')$ and $T(\mathbf{x})$ to represent the moving and fixed images, where $\mathbf{x}' = (x', y', z')$ and $\mathbf{x} = (x, y, z)$ describe 3D coordinates in the moving and fixed image domains (\mathbb{D}_S and \mathbb{D}_T , respectively), and $S(\mathbf{x}')$ and $T(\mathbf{x})$ are the intensities of each image at those coordinates. The primary goal of image registration is to estimate a transformation $\mathbf{v} : \mathbb{D}_T \rightarrow \mathbb{D}_S$, which maps corresponding locations between $S(\mathbf{x}')$ and $T(\mathbf{x})$. This is generally represented as a *pullback* vector field, $\mathbf{v}(\mathbf{x})$, where the vectors are rooted in the fixed domain and point to locations in the moving domain. The field is applied to $S(\mathbf{x}')$ by pulling moving image intensities into the fixed domain. This produces the registration result, a transformed moving image, \tilde{S} , defined as

$$\tilde{S}(\mathbf{x}) = S \circ \mathbf{v}(\mathbf{x}) = S(\mathbf{v}(\mathbf{x})) , \quad \forall \mathbf{x} \in \mathbb{D}_T , \quad (1)$$

which has coordinates in the fixed domain.

The typical registration algorithm aims to find \mathbf{v} such that the images \tilde{S} and T are as similar as possible while constraining \mathbf{v} to be smooth and continuous so that the transformation is physically sensible. This can be performed by minimizing a cost function $C(\cdot, \cdot)$ that evaluates how well aligned $S \circ \mathbf{v}(\mathbf{x})$ and $T(\mathbf{x})$ are to each other, and forcing \mathbf{v} to follow a specific transformation model. Together we can describe this problem as a standard minimization problem,

$$\arg \min_{\mathbf{v}} C(S \circ \mathbf{v}, T) , \quad (2)$$

where the transformation \mathbf{v} is the parameter being optimized.

2.2 Types of Registration

Registration algorithms are generally categorized by the transformation model used to constrain \mathbf{v} and the cost function C to evaluate similarity. The optimization approach, while important, does not usually characterize the algorithm and is often chosen to best complement the other two components of the algorithm. In this section, we cover several standard models and cost functions

that are regularly used in medical imaging. However, the actual number of registration varieties in the current literature is extensive and outside the scope of this chapter. Several literature reviews on image registration exist for a more comprehensive understanding of the subject [2, 3].

2.2.1 Types of Transformation Models

Global Transformation Models

The transformation model used to constrain \mathbf{v} in the registration algorithm is generally chosen to match the problem at hand. For example, suppose we know that the moving and fixed image is of the same person, and their only difference is caused by a turn of the head in the scanner. In such a case, we would want to use a registration algorithm that restricts \mathbf{v} to only perform translations and rotations in order to limit the possible transformation to what we expect has occurred. However, if the two images are of different people, then we might consider a more fluid transformation that can nonlinearly align parts of the anatomy. Here we will discuss two main archetypes of transformation models that are regularly used in medical imaging.

One common choice for the transformation model is to represent \mathbf{v} entirely through a global transformation on the image coordinate system. Here \mathbf{v} is described by a single linear transformation matrix M and a translation vector $\mathbf{t} = (t_x, t_y, t_z)$:

$$\mathbf{v}(\mathbf{x}) = M\mathbf{x} + \mathbf{t}. \quad (3)$$

The transformation matrix M determines the restrictiveness of the model, which is often referred to as the model's *degrees of freedom* (dof). Algorithms that only allow translations and rotations (6 dof¹) are referred to as *rigid* registrations. In such cases, M is the product of three rotation matrices (one for each axis):

$$M_{\text{rigid}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta_x & -\sin \theta_x \\ 0 & \sin \theta_x & \cos \theta_x \end{bmatrix} \begin{bmatrix} \cos \theta_y & 0 & \sin \theta_y \\ 0 & 1 & 0 \\ -\sin \theta_y & 0 & \cos \theta_y \end{bmatrix} \begin{bmatrix} \cos \theta_z & -\sin \theta_z & 0 \\ \sin \theta_z & \cos \theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (4)$$

where θ_x , θ_y , and θ_z determine the amount of rotation around each axis. If global scaling is also allowed (7 dof in total), then the algorithm becomes a *similarity* registration, and M_{rigid} is multiplied with an additional scaling matrix:

$$M_{\text{similarity}} = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & s \end{bmatrix} M_{\text{rigid}}, \quad (5)$$

¹ Here, dof are given for the 3D case since the vast majority of medical images are 3D.

where s determines the amount of scaling. Finally, adding individual scaling and shearing (12 dof in total) allows for an *affine* registration. Here the scaling matrix is modified to have independent terms s_x , s_y , and s_z for each axis, and a shear matrix is included in the product:

$$M_{\text{affine}} = \begin{bmatrix} 1 & b_{xy} & b_{xz} \\ b_{yx} & 1 & b_{yz} \\ b_{zx} & b_{zy} & 1 \end{bmatrix} \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{bmatrix} M_{\text{rigid}}, \quad (6)$$

where three pairs of shear terms describe the direction and magnitude of shearing in each axis (b_{yx} and b_{zx} for the x-axis; b_{xy} and b_{zy} for the y-axis; b_{xz} and b_{yz} for the z-axis).

The main application of these models is to account for registration problems where the moving and fixed images differ by very limited transformations. Rigid registration is regularly used to align images of the same subject, allowing for more accurate longitudinal analysis. It is also applied to images from different subjects to remove global misalignment, such as movement or shifts in position while still maintaining the physical structure in the images. Similarity and affine registrations are used when the images are expected to have differences in size or large regional transformations. In medical imaging, they offer a way to normalize different subjects in order to remove effects that are often considered unrelated to the disease being studied, such as the size of the head. In addition, affine registrations can be used to provide an initialization for more fluid registrations by removing large sweeping differences, and allowing the subsequent algorithm to focus on aligning more detailed differences. Figure 3a, b provides examples of results from rigid and affine registrations between brain MRIs from two different subjects.

Deformable Model

The main disadvantage of using only a transformation matrix to represent \mathbf{v} is its inability to account for local differences between the moving and fixed images. To perform such alignments, a *deformable* registration is necessary, where the transformation is individually defined at each point in the image using a vector field:

$$\mathbf{v}(\mathbf{x}) = \mathbf{x} + \mathbf{u}(\mathbf{x}). \quad (7)$$

The vector field \mathbf{u} is referred to as a *displacement field* and is generally restricted to be smooth and continuous to ensure the overall deformation is regularized so that the object is transformed in a physically sensible way.

Deformable registration can be loosely divided between algorithms that use parametric or nonparametric transformation models to represent \mathbf{v} . Parametric registrations use a set number of parameters to control basis functions, such as splines [4] or radial basis

functions [5], to construct and interpolate \mathbf{v} . The algorithm optimizes these parameters to find the best \mathbf{v} that minimizes the cost function. The transformations found under these models are often smooth and continuous by construction due to the basis functions used.

Nonparametric registrations are generally designed to create transformations that resemble physical motions such as elasticity [6], viscosity [7], diffusion [8], and diffeomorphism [9]. Rather than optimizing a set of parameters, the algorithm evolves the transformation at every iteration using forces imposed by the model. The strength and direction of these forces are determined by the cost function chosen and the constraints of the physical motion being modeled.

The primary application of deformable registration is to compute and align detailed correspondences between the moving and fixed images. This allows such registrations to be better suited for information transfer tasks, such as deforming anatomical labels in the moving image to match and label the same structures in the fixed image, and providing an initialization using various atlases and priors. In addition, the displacement field learned in the registration represents relative spatial change between correspondences in the moving and fixed image. Hence, it can be used to analyze morphology and shape differences between individuals [10, 11]. Figure 3c shows an example of a deformable registration performed using an adaptive bases algorithm after an affine alignment. Compared to the affine result, we see that the individual structures within the brain are now locally better aligned to match the same structures in the target brain.

2.2.2 Types of Cost Functions

The purpose of the similarity cost function is to quantify how closely aligned the transformed moving image and fixed images are to each other. Since it drives the optimization of the transformation model, the characteristics of the cost function determine what kind of images can be aligned, the degree of accuracy, and the ease of optimization. In this section, we will mainly discuss the three most popular intensity-based cost functions, which are available in most algorithms. Naturally, a large number of cost functions have been proposed in the literature, and a more complete list can be found here [2].

Sum of Square Differences.

Sum of square differences (SSD), or equivalently mean squared error (MSE), between image intensities is one of the most basic and earliest cost functions used for evaluating the similarity between two images. It consists simply of subtracting the intensity difference at each voxel between two images, squaring the difference, and then summing across all the voxels in the entire image. This can be described using

$$\mathcal{C}_{\text{SSD}}(\mathcal{T}, \tilde{\mathcal{S}}) = \sum_{\mathbf{x} \in \mathbb{D}_{\mathcal{T}}} (\mathcal{T}(\mathbf{x}) - \tilde{\mathcal{S}}(\mathbf{x}))^2. \quad (8)$$

The advantage of SSD is that it is computationally efficient, requiring only roughly three or four operations per voxel. In addition, it is very localized, since each voxel between the moving and fixed pair is calculated independently and then summed. This allows non-overlapping regions of the image to be calculated and optimized in parallel. In addition, this provides high local acuity, which allows small spatial differences between the images to be resolved by the cost function.

The main drawback of using SSD is that it is highly dependent on the absolute intensity values in the image. If correspondences in two images do not have exactly the same intensity range, the cost function will fail to register them correctly. As a result, SSD is very susceptible to errors in the presence of artifacts, intensity shifts, and partial voluming in the images.

Normalized Cross Correlation

The cross correlation (CC) function is a concept borrowed from signal processing theory for comparing the similarity between waveforms. It requires vectorizing the image (reshaping the 3D image grid into a single vector), subtracting the mean of each image, and then computing the dot product between the image vectors. The value is then divided by the magnitude of both mean subtracted vectors. This can be described by

$$\mathcal{C}_{\text{CC}}(\mathcal{T}, \tilde{\mathcal{S}}) = \left\langle \frac{(\mathcal{T} - \mu_{\mathcal{T}})}{\|\mathcal{T} - \mu_{\mathcal{T}}\|}, \frac{(\tilde{\mathcal{S}} - \mu_{\tilde{\mathcal{S}}})}{\|\tilde{\mathcal{S}} - \mu_{\tilde{\mathcal{S}}}\|} \right\rangle \quad (9)$$

$$= \frac{\sum_{\mathbf{x} \in \mathbb{D}_{\mathcal{T}}} ((\mathcal{T}(\mathbf{x}) - \mu_{\mathcal{T}})(\tilde{\mathcal{S}}(\mathbf{x}) - \mu_{\tilde{\mathcal{S}}}))}{\|\tilde{\mathcal{S}} - \mu_{\tilde{\mathcal{S}}}\| \|\mathcal{T} - \mu_{\mathcal{T}}\|}, \quad (10)$$

where $\mu_{\mathcal{T}}$ and $\mu_{\tilde{\mathcal{S}}}$ are the mean intensities of each image, and $\|\cdot\|$ indicate the ℓ_2 norm of the vectorized image intensities.

The primary advantage of CC over SSD is that it is robust to relative intensity shifts in the image, while SSD is not. This is due to the normalization using the image mean and magnitude, and the reliance on multiplication of voxel pairs instead of absolute differences. In the absence of an intensity shift, NCC can be shown to be equivalent to SSD as a cost function for optimization.

The drawback of CC is that both the mean and magnitude require a calculation over the entire image; hence, NCC loses much of the parallelization potential of SSD. In addition, the gradient on the function is more complicated to evaluate, which makes it a more difficult problem to optimize.

Mutual Information

Mutual information (MI) is a probabilistic measure of similarity derived from information theory. Using mutual information for image registration was originally presented in [12], and since then, it has become one of the most widely used registration cost functions [3]. Its success largely comes from its probabilistic nature, which gives it robustness to noise and shifts in intensity. In addition, the measure avoids evaluating direct intensity differences and instead looks at how the intensities between the two images are interdependent. This makes it a very robust measure for evaluating similarity between images with different modalities.

Mutual information is described from an information theory perspective. Hence, we start with a discrete random variable \mathcal{A} , with $P_{\mathcal{A}}(\alpha)$ representing the probability of the value α occurring in \mathcal{A} . The Shannon entropy [13] of this variable is defined by

$$H(\mathcal{A}) = - \sum_{\alpha} P_{\mathcal{A}}(\alpha) \log(P_{\mathcal{A}}(\alpha)) . \quad (11)$$

If the random variable represents image intensity values, then this entropy measures how well a given intensity value in the image can be predicted. Similarly, for a second random variable \mathcal{B} and joint probability distribution $P_{\mathcal{A},\mathcal{B}}(\alpha, b)$, the joint entropy is

$$H(\mathcal{A}, \mathcal{B}) = - \sum_{\alpha, b} P_{\mathcal{A},\mathcal{B}}(\alpha, b) \log(P_{\mathcal{A},\mathcal{B}}(\alpha, b)) , \quad (12)$$

which represents how well a given pair of intensity value in the images can be predicted. Using these terms, the mutual information is given by

$$\text{MI}(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A}, \mathcal{B}) , \quad (13)$$

which becomes

$$C_{\text{MI}}(\mathcal{T}, \tilde{\mathcal{S}}) = - (H(\mathcal{T}) + H(\tilde{\mathcal{S}}) - H(\mathcal{T}, \tilde{\mathcal{S}})) , \quad (14)$$

within the context of our registration problem. Since MI increases when the images are more similar, we negate the measure in order to fit our minimization framework.

Intuitively, mutual information describes how dependent the intensities in one image are on the other. We see that, when the images are entirely independent, the joint entropy becomes the sum of the individual entropies and the mutual information is zero. On the other hand, when the images are entirely dependent (i.e., \mathcal{v} maps \mathcal{S} exactly to \mathcal{T}), then the joint entropy becomes the entropy of the fixed image and the mutual information is maximized. In practice, the entropy and joint entropies are calculated empirically from histograms (and joint histograms) of the intensities in the images.

Since the range of entropy is sensitive to the size of the image, it is common to use a normalized variant of the measure called normalized mutual information (NMI) [14]:

$$\text{NMI}(\mathcal{T}, \tilde{\mathcal{S}}) = \frac{H(\mathcal{T}) + H(\tilde{\mathcal{S}})}{H(\mathcal{T}, \tilde{\mathcal{S}})} . \quad (15)$$

We see that this measure ranges from one to two, where two indicates a perfect alignment. Hence, we must again negate the measure when using it as a cost function to fit our minimization framework.

The main drawback of mutual information comes from its probabilistic nature. The measure relies on an accurate estimate of the probability density of the image intensities. As a result, its effectiveness decreases significantly when working with small regions within the image, where there is not enough intensity samples to accurately estimate such densities. Likewise, the measure is ineffective when facing areas of the image that have poor statistical consistency or lack clear structure [15]. Examples of this include cases where there is overwhelming noise or conversely, when the area has very homogeneous intensities and provides very little information. As a result, mutual information must be calculated over a relatively large region of the image, which reduces the measure's local acuity and diminishes its ability to handle small changes between the moving and fixed images. Lastly, as mentioned before, mutual information is almost entirely calculated from counts of intensity pairs, where the actual intensity value does not matter. While this is useful for addressing multimodal relationships, it also introduces inherent ambiguity into the measure. Given a moving and fixed image, their intensities can be paired in multiple ways to give the exact same mutual information after the transformation. Hence, the measure depends heavily on having a good initialization where the objects being registered are aligned well enough to give the correct intensity pairings at the start of the optimization. Otherwise, mutual information can cause the algorithm to align intensity pairs that incorrectly represent the correspondence between the images, resulting in registration errors [16].

3 Learning-Based Models for Registration

From the previous sections, we can see that there are numerous avenues where machine learning models can potentially be employed to address specific parts of the registration problem. We can build models to estimate the similarity between images, find anatomical correspondences in images, speed up the optimization, or even learn to estimate the transformations directly. As with most learning models, these techniques can be very broadly categorized into supervised and unsupervised techniques.

Supervised image registration within the context of machine learning entails utilizing sufficiently large training data sets of input

moving and fixed image pairs with their corresponding transformations. These data are used to train a model to learn those transformation parameters based on features discovered through the training process. The loss function quantifies the discrepancy between the predicted and input transformation parameters. For example, BIR-Net [17] presents a network for learning-based deformable registration using a dual supervision strategy where the loss is taken between the ground truth deformation field and the predicted field, in addition to the dissimilarity between the warped and fixed image. To prevent slow learning and overfitting, a hierarchical loss function is applied at various levels in the frontal part of the network. DeepFLASH [18] uses the fact that the entire optimization of large deformation diffeomorphic metric mappings (LDDMM) with geodesic shooting can be efficiently carried out in a low-dimensional bandlimited space. This motivates conversion of the velocity fields into the Fourier domain. However, neural networks that operate on complex values are inefficient and not straightforward. The method decomposes the registration framework into separable real and imaginary components and proposes the use of a dual-net that handles the real and imaginary parts separately.

One of the primary challenges with employing supervised models for image registration is that registration problems rarely have ground truth transformation data between the images. Beyond simple rigid transformations, it is too laborious and complex of a task to ask human graders to manually generate full 3D transforms between images. Instead, the desired transformations used in the training data are often obtained using outputs from traditional image registration algorithms or synthetically derived data sets, both of which can limit the capabilities of the model.

Given this limitation, more focus has been directed toward unsupervised learning-based registration approaches, which are more closely related to their traditional analogs in that they lack the use of input transformation data. Optimization is driven via loss functions which incorporate intensity-based similarity quantification in learning the correspondence between the fixed and moving images. This is conceptually analogous to the classic neural network example of unsupervised learning –the autoencoder (cf [19])– where differences between the input and the network-generated predicted version of the input are used to learn latent features characterizing the data. In the case of unsupervised image registration, the optimal transformation is that which maximizes the similarity cost function between the input, specifically the fixed image, and the network-generated predicted version of the input, specifically the warped moving image as determined by the concomitantly derived transform. Direct analogs to iterative methods can be seen in approaches such as [20], which presents a recursive cascade network where the moving image is warped iteratively to fit the

fixed image. Each subnetwork is implemented as a convolutional neural network which predicts the deformation field from the current warped image and the fixed image.

In the following sections, we will provide an overview of several key methodological archetypes in the advancement of image registration that has been made possible through the application of machine learning models. As with other parts of this chapter, it is outside of our scope to attempt to provide a comprehensive coverage of such a broad topic. Instead, we opt to lean toward more contemporary deep neural network-driven approaches, which have arisen from recent widespread adoption of deep learning models in medical image analysis. However, we encourage interested readers to explore several published review articles that can provide a more historical survey of this topic [2, 21].

3.1 Feature Extraction

Much of the early work incorporating machine learning into solving image registration problems involved the detection of corresponding features and then using that information to determine the correspondence relationship between spatial domains. These included training models to find key landmarks [22] or segmentation of structures [23], and fitting established transformations models to provide a full transformation between the images. Unsurprisingly, adaptions of these ideas carried through to deep learning approaches. For example, at the start of the current era of deep learning in image-related research, the authors of [24] proposed point correspondence detection using multiple feed-forward neural networks, each of which is trained to detect a single feature. These neural networks are relatively simple consisting of two hidden layers each with 60 neurons where the output is a probability of it containing a specific feature at the center of a small image neighborhood. These detected point correspondences are then used to estimate the total affine transformation with the RANSAC algorithm [25]. Similarly, *DeepFlow* [26] uses CNNs to detect matching features (called *deep matching*) which are then used as additional information in the large displacement optical flow framework [27]. A relatively small architecture, consisting of six layers, is used to detect features at different convolution sizes which are then matched across scales. Two algorithms for more traditional computer vision applications are proposed in [28] and [29] where both are based on the VGG architecture [30] for 2D homography estimation. The former framework includes both a regression network for determining corner correspondence and a classification network for providing confidence estimates of those predictions. The work in [29], which is publicly available, uses image patch pairs in the input layer and the ℓ_1 photometric loss between them to remove the need for direct supervision. Finally, in the category of feature learning, Wu et al. use nested auto-encoders (AE) to map patchwise image content to learned feature vectors [31]. These

patches are then subsampled based on the importance criteria outlined in [32] which tends toward regions of high informational content such as edges. The AE-based feature vectors at these image patches are then used to drive a HAMMER-based registration [33] which is inherently a feature-based, traditional image registration approach.

3.2 Domain Adaptation

In contrast to detecting discrete corresponding feature points to drive the image registration, a number of learning models have been built to predict the intensity similarity between images, directly. These techniques have largely been focused on addressing intermodality alignment, which remains an open problem due to the complexities of establishing accurate correspondence when the intensities themselves do not necessarily correspond. Models have been developed to learn intermodal spatial relationships by extending traditional concepts of image similarity, such as in [34], where intermodality transformations involving CT and MRI are learned by training on the intramodality image pairs using a basic U-net architecture and incorporating a loss function combining normalized cross correlation (NCC) and explicit regularization for enforcing smoothness of the displacement field. A related idea is developed in [35] which uses labeled data and intensity information during the training phase such that only unlabeled image data is required for prediction. The latter architecture is a densely connected U-net architecture with three types of residual shortcuts [36]. For the loss function, the authors use a multiscale Dice function with an explicit regularization term for estimating both global and local transformations. Similarity functions can also be formulated directly using learning models, such as in [37] where a two-channel network is developed for input image patches (T1- and T2-weighted brain images), and likewise, the B-spline image registration algorithm developed from the Insight Toolkit [38], which leverages the output of a CNN-based similarity measure for comparison with an identical registration setup employing mutual information.

In recent years, intermodality registration has benefited from progress made in the field of *domain adaptation*, also referred to as *image synthesis* in earlier works. The general premise behind these frameworks is that learning-based models can be used to establish the latent relationship between the intensity domains between different modalities. This allows an image in one modality to be synthesized into the other modality, or alternatively both modalities can be moved into a third artificial modality that has shared features from both modalities. When applied to image registration, these synthesized modalities can then be used to convert multi-modal registration problems into mono-modal problems that can be solved by leveraging the efficiency and accuracy of mono-modal registration techniques. [39]

Of particular note in this area are methods developed around generative adversarial networks (GANs), first introduced by Goodfellow and colleagues [40], which have increasingly found traction in addressing many types of deep learning problems in the medical imaging domain [41] including image registration. GANs are a special type of network composed of two adversarial subnetworks known as the *generator* (usually characterized by deconvolutional layers) and the *discriminator* (usually a CNN). These work in a minimax fashion to learn data distributions in the absence of extensive sample data. Seeded with a random noise image (e.g., sampled from a uniform or Gaussian distribution), the generator produces synthetic images which are then evaluated by the discriminator as belonging either to the true or synthetic data distributions in terms of some probability scalar value. This back-and-forth results in a generator network which continually improves its ability to produce data that more closely resembles the true distribution while simultaneously enhancing the discriminator’s ability to judge between true and synthetic data sets. Since the original “vanilla” GAN paper, the number of proposed GAN extensions has exploded in the literature. Initial extensions included architectural modifications for improved stability in training which have since become standard (e.g., deep convolutional GANs [42]). Please refer to Chap. 5 for a more extensive coverage of GANs.

In order to constrain the mapping between moving and fixed images, the GAN-based approach outlined in [43] combines a content loss term (which includes subterms for normalized mutual information, structural similarity [44], and a VGG-based filter feature ℓ_2 -norm between the two images) with a “cyclical” adversarial loss. This is constructed in the style of [45] who proposed this GAN extension, CycleGAN, to ensure that the normally under-constrained forward intensity mapping is consistent with a similarly generated inverse mapping for “image-to-image translation” (e.g., converting a Monet painting to a realistic photo or rendering a winter nature scene as its summer analog). However, in this case, the cyclical aspect is to ensure a regularized field through forward and inverse displacement consistency.

The work of [46] employs discriminator training between finite-element modeling and generated displacements for the prostate and surrounding tissues to regularize the predicted displacement fields. The generator loss employs the weakly supervised learning method proposed by the same authors in [47] whereby anatomical labels are used to drive registration during training only. The generator is constructed from an encoder/decoder architecture based on ResNet blocks [36]. The prediction framework includes both localized tissue deformation and the linear coordinate system changes associated with the ultrasound imaging acquisition.

In [48], the discriminator loss is based on quantification of how well two images are aligned where the negative cases derive from the registration generator and the positive cases consist of identical images (plus small perturbations). Explicit regularization is added to the total loss for the registration network which consists of a U-net type architecture that extracts two 3D image patches as input and produces a patchwise displacement field. The discriminator network takes an image pair as input and outputs the similarity probability.

3.3 Transformation Learning

Many of the methods described so far have been centered around using learning models to establish spatial correspondences between images, and then fitting traditional transformation models to align the images. An alternative approach is to directly learn and predict the transformation between images. Earlier work [49] employed CNN-based regression for estimation of 2D/3D rigid image alignment of 3D X-ray attenuation maps derived from CT and corresponding 2D digitally reconstructed (DRR) X-ray images. The transformation space is partitioned into distinct zones where each zone corresponds to a CNN-based regressor which learns transformation parameters in a hierarchical fashion. The loss function is the mean squared error on the transformation parameters.

A novel deep learning perspective was given in [50] where displacement fields are assumed to form low-dimensional manifolds and are represented in the proposed fully connected network as low-dimensional vectors. From the input vector, the network generates a 2D displacement field used to warp the moving image using bilinear interpolation. The absolute intensity difference is used to optimize the parameters of network and latent vectors. Instead of explicit regularization of the displacement field, the sum of squares of the network weights is included with the intensity error term in the loss function. Instead of training with a loss function based on similarity measures between fixed and moving images, the works of [51, 52] formulate the loss in terms of the squared difference between ground truth and predicted transformation parameters. In terms of network architecture, [51] employs a variant of U-net for training/prediction based on reference deformations provided by registration of previously segmented ROIs for cardiac matching where priority is alignment of the epicardium and endocardium. Displacement fields are parameterized by stationary velocity fields [53]. In contrast, [52] uses a smaller version of the VGG architecture to learn the parameters of a $6 \times 6 \times 6$ thin-plate spline grid.

In 2015, Jaderberg and his fellow co-authors described a powerful new module, known as the spatial transformer network (STN) [54]² which features prominently now in many contemporary deep

² Note that these networks are different from transformers and visual transformers described in Chap. 6.

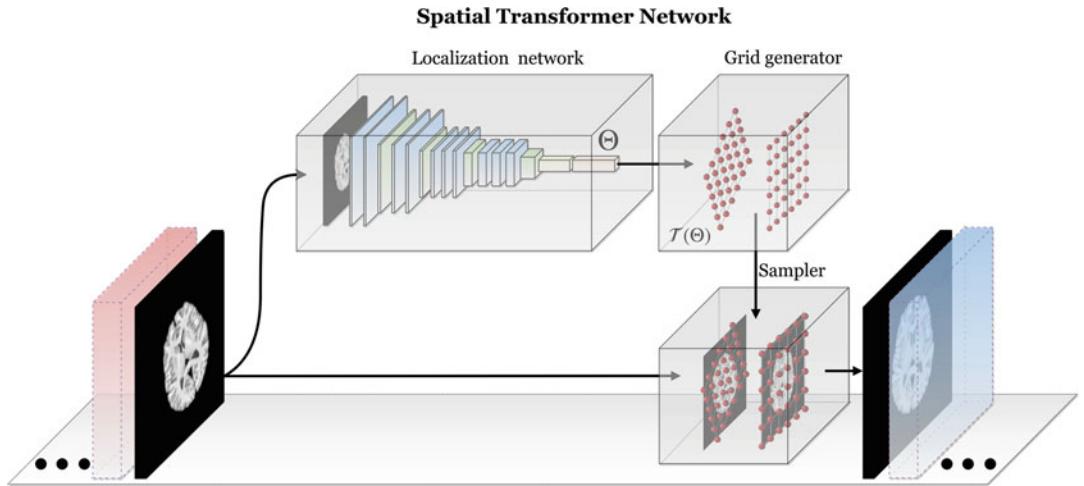


Fig. 4 Diagrammatic illustration of the spatial transformer network. The STN can be placed anywhere within a CNN to provide spatial invariance for the input feature map. Core components include the localization network used to learn/predict the parameters which transform the input feature map. The transformed output feature map is generated with the grid generator and sampler. ©2019 Elsevier. Reprinted, with permission, from [21]

learning-based registration approaches. Generally, STNs enhance CNNs by permitting a flexibility which allows for an explicit spatial invariance that goes beyond the implicitly limited translational invariance associated with the architecture's pooling layers. In many image-based tasks (e.g., localization or segmentation), designing an algorithm that can account for possible pose or geometric variation of the object(s) of interest within the image is crucial for maximizing performance. The STN is a fully differentiable layer which can be inserted anywhere in the CNN to learn the parameters of the transformation of the input feature map (not necessarily an image) which renders the output in such a way so as to optimize the network based on the specified loss function. The added flexibility and the fact that there is no manual supervision or special handling required make this module an essential addition for any CNN-based toolkit.

An STN comprises three principal components: (1) a localization network, (2) a grid generator, and (3) a sampler (*see* Fig. 4). The localization network uses the input feature map to learn/regress the transformation parameters which optimize a specified loss function. In many examples provided, this amounts to transforming the input feature map to a quasi-canonical configuration. The actual architecture of the localization network is fairly flexible, and any conventional architecture, such as a fully connected network (FCN), is suitable as long as the output maps to the continuous estimate of the transformation parameters. These transformation parameters are then applied to the output of the grid generator which are simply the regular coordinates of the input

image (or some normalized version thereof). The sampler, or interpolator, is used to map the transformed input feature map to the coordinates of the output feature map.

Since Jaderberg’s original STN formulation, extensions have been proposed such as the inverse compositional STN (IC-STN) [55] and the diffeomorphic transformer network [56]. Two issues with the STN include the following: (1) potential boundary effects in which learned transforms require sampling outside the boundary of the input image which can cause potential learning errors for subsequent layers and (2) the single-shot estimate of the learned transform which can compromise accuracy for large transformation distances. The IC-STN addresses both of these issues by (1) propagating transformation parameters instead of propagating warped input feature maps until the final transformation layer and (2) recurrent usage of the localization network for inferring transform compositions in the spirit of the inverse compositional Lucas-Kanade algorithm [57].

Although discussion of transform generalizability was included in the original STN paper [54], discussion was limited to affine, attention (scaling + translation), and thin-plate spline transforms which all comply with the requirement of differentiability. This work was extended to diffeomorphic transforms in [56]. The computational load associated with generating traditional diffeomorphisms through velocity field integration [58] motivated the use of continuous piecewise affine-based (CPAB) transformations [59]. The CPAB approach utilizes a tesselation of the image domain which translates into faster and more accurate generation of the resulting diffeomorphism. Although this does constrain the flexibility of the final transformation, the framework provides an efficient compromise for use in deep learning architectures. Analogous to traditional image registration, the deep diffeomorphic transformer layer can be placed in serial following an affine-based STN layer for a global-to-local total transformation estimation. This is demonstrated in the experiments reported in [56].

The development of the STN has led to a number of notable generalized deep learning-based registration approaches. *Voxel-Morph*, first presented in [60], incorporates a U-net architecture with a STN where the input layer consists of the concatenated full fixed and moving image volumes resized and cropped to $160 \times 192 \times 224$ voxels. The output consists of the voxelwise displacement field of the same size as the input (times three—one for each vector component). The loss function for training combines cross correlation and a diffusion regularizer on the spatial gradients of the displacement field. This was extended to a generative approach in [61] to yield diffeomorphic transformations based on SVFs [53] using novel scaling and squaring network layers. The U-net architecture is used to estimate the distribution parameters of the velocity fields encapsulated by training data. A new imaging

pair can then be registered by sampling from this learned distribution, computing the resulting diffeomorphic transformation, and then warping the moving image. The underlying code has been made publically available which has facilitated independent evaluations such as [62] to compare performance with traditional algorithms (i.e., IRTK [63], AIR [64], Elastix [65], ANTs [66], and NiftyReg [67]). Other variations include CycleMorph [68], which uses a cycle-consistency objective to learn to produce the original image from the deformed image conditioned on the transformation. This prevents degeneracies in the learned registration fields and demonstrates the potential to preserve topologies by inducing cycle consistency on the images. Another generative image registration approach is that of [69] which uses a conditional variational autoencoder [70], an extension of the variational autoencoder [71] which permits incorporation of additional information for latent inference modeling. This multi-scale generative framework encodes the SVFs which are ultimately converted to the total transformation field in a similar fashion as [61].

3.4 Optimization and Equation Solving

A current limitation of traditional registration techniques is the computation cost associated with finding an iterative solution. Most existing registration methods do not scale linearly with image size; thus, as advancements in medical imaging lead to increasingly higher resolution data, the time scale to operate registration techniques can expand to hours, and possibly days, per registration. While not specific to image registration, one area of research that can help address this is the application of learning models to replace classic optimization and equation solving techniques. These can lead to dramatic speed up of existing registration techniques while maintaining the same transformation models. Examples of advancements in this area include the use of learning-based ODE solutions to perform diffeomorphic registration [72] and the use of deep learning to initialize classical optimization approaches, such as Newton's method [73].

4 Registration in the Study of Brain Disorders

This final section will explore how learning-based models have impacted several primary applications of image registration, particularly for the study of diseases. As before, this discussion is far from comprehensive, but more to demonstrate current trends in using machine learning models to advance common areas of registration-driven image analysis.

4.1 Spatial Normalization and Atlasing

Normative and disease-specific atlases play an important role in the characterization of a disease. By registering images from different subjects into a common atlas space (i.e., *spatial normalization*), we can remove typical variability between subjects, such as brain size, to allow for more sensitive detection of disease-driven differences between subjects. Learning-based registration can enable higher throughput registration during atlas construction [74], thus allowing more subjects to be included into the atlas and better encompassing the variability within a cohort. Various models have been proposed to embed these advantages directly into the network, such as [75], which uses a joint learning framework where image attributes are used to learn conditional templates, and an efficient deformation to these templates is jointly learned. In addition, learning models have been used to provide priors for the atlas [76] and establish groupwise correspondence within a cohort [77].

4.2 Label Transfer

As described in earlier sections, establishing correspondences between images via image registration allows for the transfer of spatially embedded data, such as structural annotations and segmentations, between different images and subjects. This method, colloquially referred to as *label transfer*, allows for automatic identification of anatomy in the image that may be relevant to a disease. While a natural application of learning models for label transfer is to simply replace traditional registration approaches with learning-based ones, there has also been more sophisticated integration of machine learning into these frameworks. Popular among these are joint techniques that aim to integrate and solve for both the segmentation and registration problem simultaneously in the same framework [78, 79]. For example, LT-Net [80] learns a multi-atlas registration using cycle consistency and a LSGAN objective [81] to discriminate synthesized images from real ones. Cycle consistency is applied in the image space (between the true atlas and the reconstructed atlas), the transformation space (a voxel warped from the forward transformation composed with the reversed transformation would end up in its starting point), and the segmentation label space. Learning models have also been shown to be effective for correcting systematic errors in both the registration and segmentation parts of the framework [82]. Other models have been proposed for replacing non-registration parts of the standard multi-atlas label transfer framework, such as the voting scheme [83].

4.3 Morphometry

Voxel-based [84] and tensor-based [85] morphometry is the analysis of the transformation result from an image registration to study the shape and structural characteristics of a disease. In these approaches, a disease cohort is spatially normalized into a common space and the warped images and resulting deformation fields from each registration are statistically compared on a voxel level to reveal

morphological characteristics in the cohort. Machine learning models offer new ways to analyze the resulting morphology, such as integrating them as part of a multivariate biomarker framework to detect a disease [86, 87].

5 Conclusion

Image registration is a core pillar of modern-day image analysis, allowing for the alignment and transfer of spatial information between subjects and imaging modalities. Learning-based models have marked improvements on core aspects of image registration, ranging from more accurate feature detection, to better intensity correspondences, particularly across modalities, to improving the speed and accuracy of the alignment.

Acknowledgements

The authors wish to acknowledge the staff and researchers at the Penn Image Computing and Science Laboratory (PICSL) for their support and expertise.

References

1. Anuta PE (1970) Spatial registration of multispectral and multitemporal digital imagery using fast Fourier transform techniques. *IEEE Trans Geosci Electron* 8(4):353–368
2. Sotiras A, Davatzikos C, Paragios N (2013) Deformable medical image registration: a survey. *IEEE Trans Med Imaging* 32(7):1153–1190
3. Pluim JP, Maintz JA, Viergever MA (2003) Mutual-information-based registration of medical images: a survey. *IEEE Trans Med Imaging* 22(8):986–1004
4. Bookstein FL (1989) Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.* 11(6):567–585
5. Rohde GK, Aldroubi A, Dawant BM (2003) The adaptive bases algorithm for intensity-based nonrigid image registration. *IEEE Trans Med Imaging* 22(11):1470–1479
6. Gee JC, Bajcsy RK (1998) Elastic matching: continuum mechanical and probabilistic analysis. *Brain Warping* 2
7. Christensen GE, Rabbitt RD, Miller MI (1996) Deformable templates using large deformation kinematics. *IEEE Trans. Image Process.* 5(10):1435–1447
8. Thirion JP (1998) Image matching as a diffusion process: an analogy with Maxwell’s demons. *Med Image Anal* 2(3):243–260
9. Beg MF, Miller MI, Trouvé A, Younes L (2005) Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int J Comput Vis* 61(2):139–157
10. Ashburner J, Friston KJ (2000) Voxel-based morphometry—the methods. *NeuroImage* 11(6):805–821
11. Davatzikos C, Genc A, Xu D, Resnick SM (2001) Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. *NeuroImage* 14(6):1361–1369
12. Wells III WM, Viola P, Atsumi H, Nakajima S, Kikinis R (1996) Multi-modal volume registration by maximization of mutual information. *Med Image Anal* 1(1):35–51
13. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423
14. Studholme C, Hill DL, Hawkes DJ (1999) An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recogn* 32(1):71–86
15. Andronache A, von Siebenthal M, Székely G, Cattin P (2008) Non-rigid registration of

- multi-modal images using both mutual information and cross-correlation. *Med Image Anal* 12(1):3–15
16. Maes F, Vandermeulen D, Suetens P (2003) Medical image registration using mutual information. *Proc IEEE* 91(10):1699–1722
 17. Fan J, Cao X, Yap PT, Shen D (2019) BIRNet: brain image registration using dual-supervised fully convolutional networks. *Med Image Anal* 54:193–206
 18. Wang J, Zhang M (2020) Deepflash: an efficient network for learning-based medical image registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4444–4452
 19. Hinton GE, Zemel RS (1994) Autoencoders, minimum description length and Helmholtz free energy. In: Cowan JD, Tesauro G, Alspector J (eds) *Advances in neural information processing systems*. Morgan-Kaufmann, Burlington, pp 3–10
 20. Zhao S, Dong Y, Chang EIC, Xu Y (2019) Recursive cascaded networks for unsupervised medical image registration. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV)
 21. Tustison NJ, Avants BB, Gee JC (2019) Learning image-based spatial transformations via convolutional neural networks: a review. *Magn Reson Imaging* 64:142–153
 22. Ozysal M, Calonder M, Lepetit V, Fua P (2009) Fast keypoint recognition using random ferns. *IEEE Trans Pattern Anal Mach Intell* 32(3):448–461
 23. Powell S, Magnotta VA, Johnson H, Jammalamadaka VK, Pierson R, Andreasen NC (2008) Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. *NeuroImage* 39(1): 238–247
 24. Sergeev S, Zhao Y, Linguraru MG, Okada K (2012) Medical image registration using machine learning-based interest point detector. In: Proceedings of the SPIE
 25. Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm ACM* 24(6): 381–395
 26. Weinzaepfel P, Revaud J, Harchaoui Z, Schmid C (2013) Deepflow: large displacement optical flow with deep matching. In: Proceedings of the IEEE international conference on computer vision, pp 1385–1392. <https://doi.org/10.1109/ICCV.2013.175>
 27. Brox T, Malik J (2011) Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Trans Pattern Anal Mach Intell* 33(3):500–513. <https://doi.org/10.1109/TPAMI.2010.143>
 28. DeTone D, Malisiewicz T, Rabinovich A (2016) Deep image homography estimation. arXiv:160603798
 29. Nguyen T, Chen SW, Shivakumar SS, Taylor CJ, Kumar V (2018) Unsupervised deep homography: a fast and robust homography estimation model. In: Proceedings of IEEE robotics and automation letters
 30. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556, 1409.1556
 31. Wu G, Kim M, Wang Q, Munsell BC, Shen D (2016) Scalable high-performance image registration framework by unsupervised deep feature representations learning. *IEEE Trans Biomed Eng* 63(7):1505–1516. <https://doi.org/10.1109/TBME.2015.2496253>
 32. Wang Q, Wu G, Yap PT, Shen D (2010) Attribute vector guided groupwise registration. *NeuroImage* 50(4):1485–1496. <https://doi.org/10.1016/j.neuroimage.2010.01.040>
 33. Shen D, Davatzikos C (2002) Hammer: hierarchical attribute matching mechanism for elastic registration. *IEEE Trans Med Imaging* 21(11):1421–1439. <https://doi.org/10.1109/TMI.2002.803111>
 34. Cao X, Yang J, Zhang J, Nie D, Kim MJ, Wang Q, Shen D (2017) Deformable image registration based on similarity-steered cnn regression. In: Proceedings of the international conference on medical image computing and computer-assisted intervention 10433:300–308. https://doi.org/10.1007/978-3-319-66182-7_35
 35. Hu Y, Modat M, Gibson E, Li W, Ghavami N, Bonmati E, Wang G, Bandula S, Moore CM, Emberton M, Ourselin S, Noble JA, Barratt DC, Vercauteren T (2018) Weakly-supervised convolutional neural networks for multimodal image registration. *Med Image Anal* 49:1–13. <https://doi.org/10.1016/j.media.2018.07.002>
 36. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. CoRR abs/1512.03385. <http://arxiv.org/abs/1512.03385>, 1512.03385
 37. Simonovsky M, Gutierrez-Becker B, Mateus D, Navab N, Komodakis N (2016) A deep metric for multimodal registration. In: Proceedings of the international conference on medical image computing and computer-assisted intervention

38. Yoo TS, Metaxas DN (2005) Open science—combining open data and open source software: medical image analysis with the insight toolkit. *Med Image Anal* 9(6):503–6. <https://doi.org/10.1016/j.media.2005.04.008>
39. Chen M, Carass A, Jog A, Lee J, Roy S, Prince JL (2017) Cross contrast multi-channel image registration using image synthesis for mr brain images. *Med Image Anal* 36:2–14
40. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems
41. Yi X, Walia E, Babyn P (2018) Generative adversarial network in medical imaging: a review. Preprint
42. Radford A, Metz L, Chintala S (2016) Unsupervised representation learning with deep convolutional generative adversarial networks. In: Proceedings of the international conference on learning representations
43. Mahapatra D, Antony B, Sedai S, Garnavi R (2018) Deformable medical image registration using generative adversarial networks. In: Proceedings of IEEE 15th international symposium on biomedical imaging
44. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 13(4):600–612
45. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE international conference on computer vision
46. Hu Y, Gibson E, Ghavami N, Bonmati E, Moore CM, Emberton M, Vercauteren T, Noble JA, Barratt DC (2018) Adversarial deformation regularization for training image registration neural networks. In: Proceedings of the international conference on medical image computing and computer-assisted intervention
47. Hu Y, Modat M, Gibson E, Ghavami N, Bonmati E, Moore CM, Emberton M, Noble JA, Barratt DC, Vercauteren T (2018) Label-driven weakly-supervised learning for multi-modal deformable image registration. In: Proceedings of IEEE 15th international symposium on biomedical imaging
48. Fan J, Cao X, Xue Z, Yap PT, Shen D (2018) Adversarial similarity network for evaluating image alignment in deep learning based registration. In: Proceedings of the international conference on medical image computing and computer-assisted intervention
49. Miao S, Wang ZJ, Liao R (2016) A CNN regression approach for real-time 2D/3D registration. *IEEE Trans Med Imaging* 35(5):1352–1363. <https://doi.org/10.1109/TMI.2016.2521800>
50. Sheikhjafari A, Noga M, Punithakumar K, Ray N (2018) Unsupervised deformable image registration with fully connected generative neural network. In: Proceedings of medical imaging with deep learning
51. Rohé MM, Datar M, Heimann T, Sermesant M, Pennec X (2017) SVF-Net: learning deformable image registration using shape matching. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S (eds) Proceedings of the international conference on medical image computing and computer-assisted intervention. Springer International Publishing, Cham, pp 266–274
52. Eppenhof KAJ, Lafarge MW, Moeskops P, Veta M, Pluim JPW (2018) Deformable image registration using convolutional neural networks. In: Proceedings of the SPIE: medical imaging: image processing
53. Arsigny V, Commowick O, Pennec X, Ayache N (2006) A log-euclidean framework for statistics on diffeomorphisms. In: Proceedings of the international conference on medical image computing and computer-assisted intervention, vol 9, pp 924–31
54. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks. In: Neural information processing systems
55. Lin CH, Lucey S (2017) Inverse compositional spatial transformer networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition
56. Detlefsen NS, Freifeld O, Hauberg S (2018) Deep diffeomorphic transformer networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition
57. Baker S, Matthews I (2004) Lucas-kanade 20 years on: a unifying framework. *Int J Comput Vis* 56(3):221–255. <https://doi.org/10.1023/B:VISI.0000011205.11775.fd>
58. Beg MF, Miller MI, Trouvé A, Younes L (2004) Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int J Comput Vis* 61(2):139–157
59. Freifeld O, Hauberg S, Batmanghelich K, Fisher JW (2017) Transformations based on continuous piecewise-affine velocity fields. *IEEE Trans Pattern Anal Mach Intell* 39(12):2496–2509. <https://doi.org/10.1109/TPAMI.2016.2646685>

60. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV (2018) An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE conference on computer vision and pattern recognition
61. Dalca AV, Balakrishnan G, Guttag J, Sabuncu MR (2018) Unsupervised learning for fast probabilistic diffeomorphic registration. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G (eds) Proceedings of the international conference on medical image computing and computer-assisted intervention. Springer International Publishing, Cham, pp 729–738
62. Nazib A, Fookes C, Perrin D (2018) A comparative analysis of registration tools: traditional vs deep learning approach on high resolution tissue cleared data. Preprint. arXiv
63. Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ (1999) Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans Med Imaging* 18(8):712–721. <https://doi.org/10.1109/42.796284>
64. Woods RP, Mazziotta JC, Cherry SR (1993) MRI-PET registration with automated algorithm. *J Comput Assist Tomogr* 17(4): 536–546
65. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW (2010) Elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging* 29(1):196–205. <https://doi.org/10.1109/TMI.2009.2035616>
66. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC (2011) A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage* 54(3): 2033–2044. <https://doi.org/10.1016/j.neuroimage.2010.09.025>
67. Modat M, Ridgway GR, Taylor ZA, Lehmann M, Barnes J, Hawkes DJ, Fox NC, Ourselin S (2010) Fast free-form deformation using graphics processing units. *Comput Methods Prog Biomed* 98(3):278–284. <https://doi.org/10.1016/j.cmpb.2009.09.002>
68. Kim B, Kim DH, Park SH, Kim J, Lee JG, Ye JC (2021) Cyclemorph: cycle consistent unsupervised deformable image registration. *Med Image Anal* 71:102036
69. Krebs J, Mansi T, Mailhé B, Ayache N, Delingette H (2018) Unsupervised probabilistic deformation modeling for robust diffeomorphic registration. In: Proceedings of the 4th international workshop, DLMIA and 8th international workshop, ML-CDS
70. Sohn K, Lee H, Yan X (2015) Learning structured output representation using deep conditional generative models. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems 28. Curran Associates Inc., Red Hook, pp 3483–3491
71. Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: Proceedings of the 2nd international conference on learning representations (ICLR)
72. Wu Y, Jiahao TZ, Wang J, Yushkevich PA, Hsieh MA, Gee JC (2022) Nodeo: a neural ordinary differential equation based optimization framework for deformable image registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 20804–20813
73. Huang J, Wang H, Yang H (2020) Int-deep: a deep learning initialized iterative method for nonlinear problems. *J Comput Phys* 419: 109675
74. Iqbal A, Khan R, Karayannidis T (2019) Developing a brain atlas through deep learning. *Nat Mach Intell* 1(6):277–287
75. Dalca A, Rakic M, Guttag J, Sabuncu M (2019) Learning conditional deformable templates with convolutional networks. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems. Curran Associates Inc., Red Hook, vol 32. <https://proceedings.neurips.cc/paper/2019/file/bbcbff5c1f1ded46c25d28119a85c6c2-Paper.pdf>
76. Wu N, Wang J, Zhang M, Zhang G, Peng Y, Shen C (2022) Hybrid atlas building with deep registration priors. In: 2022 IEEE 19th international symposium on biomedical imaging (ISBI). IEEE, Piscataway, pp 1–5
77. Yang J, Küstner T, Hu P, Liò P, Qi H (2022) End-to-end deep learning of non-rigid group-wise registration and reconstruction of dynamic MRI. *Front Cardiovasc Med* 9
78. Sinclair M, Schuh A, Hahn K, Petersen K, Bai Y, Batten J, Schaap M, Glocker B (2022) Atlas-ISTN: joint segmentation, registration and atlas construction with image-and-spatial transformer networks. *Med Image Anal* 78: 102383
79. Xu Z, Niethammer M (2019) Deepatlas: joint semi-supervised learning of image registration and segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, Berlin, pp 420–429

80. Wang S, Cao S, Wei D, Wang R, Ma K, Wang L, Meng D, Zheng Y (2020) LT-Net: label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)
81. Mao X, Li Q, Xie H, Lau RY, Wang Z, Paul Smolley S (2017) Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2794–2802
82. Wang H, Yushkevich PA (2013) Multi-atlas segmentation with joint label fusion and corrective learning—an open source implementation. *Front Neuroinform* 7:27
83. Ding Z, Han X, Niethammer M (2019) Votenet: a deep learning label fusion method for multi-atlas segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, Berlin, pp 202–210
84. Ashburner J, Friston KJ (2000) Voxel-based morphometry—the methods. *NeuroImage* 11(6 Pt 1):805–821. <https://doi.org/10.1006/nimg.2000.0582>
85. Hua X, Leow AD, Parikshak N, Lee S, Chiang MC, Toga AW, Jack Jr CR, Weiner MW, Thompson PM, Initiative ADN, et al (2008) Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. *NeuroImage* 43(3):458–469
86. Pahuja G, Prasad B (2022) Deep learning architectures for Parkinson's disease detection by using multi-modal features. *Comput Biol Med* 105610
87. Huang H, Zheng S, Yang Z, Wu Y, Li Y, Qiu J, Cheng Y, Lin P, Lin Y, Guan J, et al (2022) Voxel-based morphometry and a deep learning model for the diagnosis of early alzheimer's disease based on cerebral gray matter changes. *Cereb Cortex* 33(3):754–763

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 15

Computer-Aided Diagnosis and Prediction in Brain Disorders

Vikram Venkatraghavan, Sebastian R. van der Voort, Daniel Bos, Marion Smits, Frederik Barkhof, Wiro J. Niessen, Stefan Klein, and Esther E. Bron

Abstract

Computer-aided methods have shown added value for diagnosing and predicting brain disorders and can thus support decision making in clinical care and treatment planning. This chapter will provide insight into the type of methods, their working, their input data –such as cognitive tests, imaging, and genetic data– and the types of output they provide. We will focus on specific use cases for diagnosis, i.e., estimating the current “condition” of the patient, such as early detection and diagnosis of dementia, differential diagnosis of brain tumors, and decision making in stroke. Regarding prediction, i.e., estimation of the future “condition” of the patient, we will zoom in on use cases such as predicting the disease course in multiple sclerosis and predicting patient outcomes after treatment in brain cancer. Furthermore, based on these use cases, we will assess the current state-of-the-art methodology and highlight current efforts on benchmarking of these methods and the importance of open science therein. Finally, we assess the current clinical impact of computer-aided methods and discuss the required next steps to increase clinical impact.

Key words Dementia, Stroke, Glioma, Cognitive impairment

1 Introduction

Computer-aided methods have major potential value for diagnosing and predicting outcomes in brain disorders such as dementia, brain cancer, and stroke. Diagnosis aims to determine the current “condition” of the patient. Prediction, or prognosis, on the other hand, aims to forecast the future “condition” of the patient. In this way, the patient’s current and future condition can be estimated in a more detailed and accurate way, which opens up possibilities for better patient care and personalized medicine, with interventions tailored to the individual patient. Moreover, diagnosis and

Authors Stefan Klein and Esther E. Bron have contributed equally to this chapter

prediction are crucial not only for decision making in clinical care and treatment planning but also for managing the expectations of patients and their caregivers. This is particularly important in brain disorders as they may strongly affect life expectancy and quality of life, as symptoms of the disorder and side effects of the treatment can have a major impact on the patient's cognitive skills, daily functioning, social interaction, and general well-being. In clinical practice, diagnosis and prediction are typically performed using multiple sources of information, such as symptomatology, medical history, cognitive tests, brain imaging, electroencephalography (EEG), magnetoencephalography (MEG), blood tests, cerebrospinal fluid (CSF) biomarkers, histopathological or molecular findings, and lifestyle and genetic risk factors. These various pieces of information are integrated by the treating clinician, often in consensus with other experts at a multidisciplinary team meeting, in order to reach a final diagnosis and/or treatment plan. The aim of computer-aided methods for diagnosis and prediction is to support this process, in order to achieve more accurate, objective, and efficient decision making.

In the literature, numerous examples of computer-aided methods for diagnosis and prediction in brain disorders can be found. Most of the state-of-the-art methods use some form of machine learning to construct a model that maps (often high-dimensional) input data to the output variable of interest. There exists a large variation in machine learning technology, types of input data, and output variables. Chapters 1–6 introduced the main machine learning technologies used for computer-aided diagnosis and prediction. These include, on the one hand, classical methods such as linear models, support vector machines, and random forests, and on the other hand, deep learning methods such as convolutional neural networks and recurrent neural networks. These methods can be implemented either as classification models (estimating discrete labels) or as regression models (estimating continuous quantities), possibly specialized for survival (or “time-to-event”) analysis. In addition, Chapter 17 highlights the category of disease progression modeling techniques, which could be considered as a specialized form of machine learning incorporating models of the disease evolution over time. Chapters 7–12 described the main types of input data used in machine learning for brain disorders: clinical evaluations, neuroimaging, EEG/MEG, genetics and omics data, electronic health records, and smartphone and sensor data. The current chapter focuses on the choice of the output variable, i.e., the diagnosis or prediction of interest (Fig. 1).

To illustrate the various ways in which machine learning could aid diagnosis and prediction, we focus on representative use cases

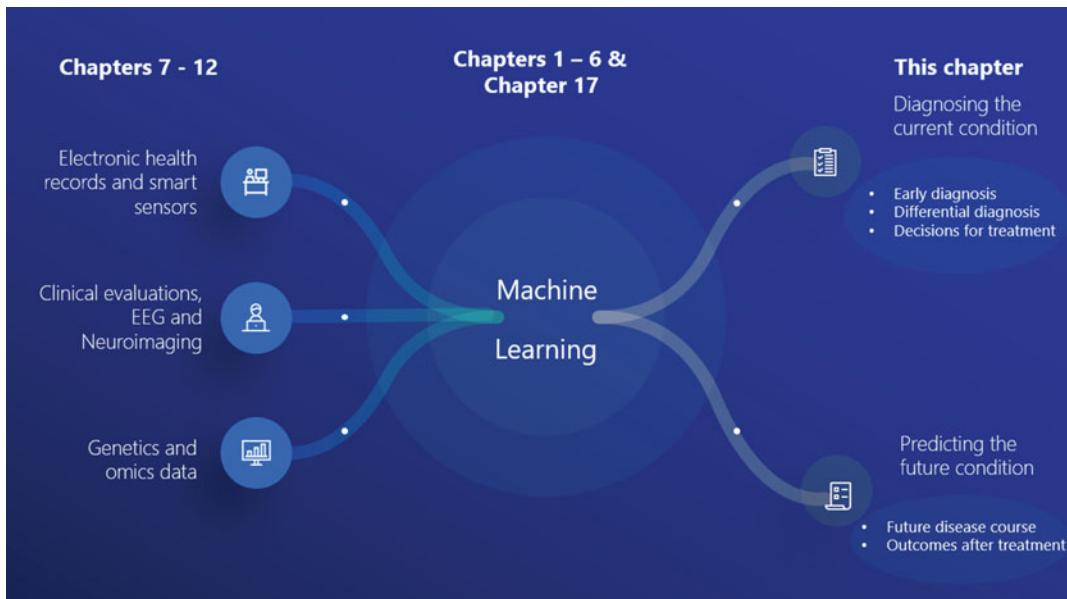


Fig. 1 Overview of the topics covered in this chapter, in the context of the other chapters in this book

organized according to the type of output. Subheading 1.1 presents diagnostic use cases, including early diagnosis, differential diagnosis, and decision making for treatment. Subheading 1.2 presents prediction use cases, including estimation of the natural disease course and prediction of patient outcomes after treatment. While the diagnostic use cases are the core of current clinical practice which could be aided by machine learning, the prediction use cases represent a potential future application. Currently, prediction is not so often made as clinicians are not yet able to make a reliable prediction in most cases. After these introductory sections, Subheading 2 provides a more comprehensive survey of the state-of-the-art methodology, and Subheading 3 analyzes the clinical impact of such methodology and suggests a roadmap for further clinical translation. Finally, Subheading 5 concludes this chapter.

1.1 Diagnosis

Diagnosis aims to determine the current “condition” of the patient to inform patient care and treatment decisions. Here, we introduce three categories of diagnostic tasks that occur in clinical practice and describe why and how computer-aided models have or could have added value.

Box 1: Diagnosis

Categories of diagnostic tasks that occur in clinical practice in which computer-aided models have or could have added value, with brain disorders for which this is relevant as examples:

- **Early diagnosis** Dementia and MS
- **Differential diagnosis** Dementia and brain cancer
- **Decision making for treatment** Stroke

Early diagnosis is highly challenging in neurodegenerative diseases such as dementia and multiple sclerosis (MS). Dementia is a clinical syndrome which can be caused by several underlying diseases, Alzheimer's disease (AD) being the most prevalent, and is estimated to affect 50 million people worldwide [4]. The mean age at dementia diagnosis is approximately 83 years [106]. MS is estimated to affect about two million people worldwide, and it primarily affects younger adults with the mean age of onset for incident MS being approximately 30 years [74]. Both for dementia and MS, establishing the diagnosis usually takes a substantial period of time after the first clinical symptoms arise [58, 139]. Early detection and accurate diagnosis is crucial for timely decision making regarding care and management of dementia symptoms, and as such can reduce healthcare costs and improve quality of life as it gives patients access to supportive therapies that help to delay institutionalization [107]. Early diagnosis of MS is important, because patients who begin treatment earlier do reap more benefit than those who start late [90]. In addition, advancing the diagnosis in time is essential to support the development of new disease-modifying treatments, since late treatment is expected to be a major factor in failure of clinical trials [88]. The clinical diagnosis of dementia is currently based on objective assessment of cognitive impairment, assessment of biomarkers [29], and evaluation of its interference with daily living [2, 42, 87, 112]. The clinical diagnosis of MS is based on frequency of relapsing inflammatory attacks, associated symptoms, and distribution of lesions on MRI [132]. For a subset of MS patients with demyelinating lesions highly suggestive of MS, termed as radiologically isolated syndrome (RIS), a separate diagnostic criteria was formed by Okuda et al. [98] to improve the diagnostic accuracy. However, objective assessment of biomarkers of the underlying processes can advance diagnosis, since symptoms are known to arise relatively late in the disease process. This holds, for example, for cognitive impairment due to dementia and physical disability or cognitive impairment due to MS [25, 40, 52]. By combining neuroimaging and other biomarkers

with machine learning based on large datasets, computer-aided diagnosis algorithms aim to facilitate medical decision support by providing a potentially more objective diagnosis than that obtained by conventional clinical criteria [63, 113]. In addition to biomarkers, machine learning based on data from remote monitoring technology, such as wearables and smart watches, is an emerging field of research aimed at detecting cognitive, behavioral, and physical symptoms in an objective way at the earliest stage possible [95, 126].

Beyond an early diagnosis, accurate identification of the underlying disease, i.e., **differential diagnosis**, is crucial for planning care and treatment decisions. For example, in dementia, the most common underlying diseases are AD, vascular cognitive impairment (VCI), dementia with Lewy bodies (DLB), and frontotemporal lobar degeneration (FTLD). Although clinical symptomatology differs between the diseases, symptoms in the early stage may be unclear and can overlap [42, 87, 112]. The current clinical criteria for AD and FTLD, for example, which entail qualitative inspection of neuroimaging, fail to accurately differentiate the two diseases [47]. Additionally, a young patient (< 65 years old) with behavioral problems could have a differential diagnosis of dementia (i.e., behavioral phenotypes of FTLD or AD) or primary psychiatric disorder, as symptomatology overlaps substantially [68]. An accurate diagnosis of primary psychiatric disorder can be informative in such patients by suggesting that progressive decline in the condition is not necessarily expected [30]. For some specific diseases, measurements of proteins causing the underlying pathology have in the last decade shown high accuracy for diagnosis of the pathology. AD is a good example with blood-based biomarkers measuring phosphorylated-Tau (P-Tau), CSF biomarkers measuring amyloid β , P-Tau and Tau, and PET imaging measuring amyloid- β and Tau. However, while highly promising, measurement of these proteins is not yet widely performed in clinical practice as blood-based biomarkers of AD are not widely available yet, CSF biomarkers require an invasive lumbar puncture, and PET imaging is too expensive and not sufficiently widely accessible to be done in each patient. Moreover, such markers of the underlying pathology are currently unavailable for other types of dementia. As an alternative, quantitative neuroimaging and other biomarkers, especially in combination with machine learning and large datasets, have shown to be beneficial in difficult cases of differential diagnosis [14, 110].

Another disorder where differential diagnosis is crucial is brain cancer. Diagnosis of brain tumors typically starts with the analysis of MRI brain data. A first diagnostic task is to differentiate between primary and secondary lesions. Primary lesions are tumors that originated from healthy brain cells, with glioma being the most common primary brain tumor type. Secondary lesions are metastases from tumors located elsewhere in the body, which may trigger

very different care and treatment paths. Also the distinction between glioma and other less common malignant primary lesions such as lymphoma is relevant. Whereas neuroradiologists are trained to differentiate these different types of lesions, the large variation in appearance of tumors induces uncertainty in the differential diagnosis. Machine learning has been shown to be able to distinguish glioma from metastasis [20] and lymphoma [86] based on quantitative analysis of brain MRI, and may thus be used as a “second” reader supporting the radiologists. Once a diagnosis of cancer is established, a second task in differential diagnosis is the further subtyping of the lesion. While glioma is one of the deadliest forms of cancer [97], there exist large differences in survival and treatment response between patients. These differences can be attributed to the glioma’s genetic and histological features, in particular the isocitrate dehydrogenase (IDH) mutation status, the 1p19q co-deletion status, MGMT promoter methylation status, and the tumor grade [28, 31, 38]. These insights have led to classification guidelines by the World Health Organization (WHO) [77]. In current clinical practice, these genetic and histological features are determined from tumor tissue after resection. However, there has been an increasing interest in complementary non-invasive alternatives that can provide the genetic and histological information before resection [10, 152]. Also here, neuroradiologists can be trained to visually distinguish the subtypes based on MRI [26, 128], but uncertainty often remains and the inherent subjectivity associated with visual inspection of subtle differences in appearance, by radiologists with varying levels of expertise, is undesirable. A large body of research has therefore focused on development of machine learning approaches to support MRI-based determination of genetic and histological features of glioma [41, 65, 122, 127].

The third diagnostic task we address is **decision making for treatment**. This is relevant when multiple therapeutic options are available, such as for patients with stroke. Multiple treatment options for stroke exist such as thrombolytic medication and endovascular clot retrieval (mechanical thrombectomy). Since depending on the situation different treatments or their combination may be optimal, and since the costs per patient are rising, there is a real and urgent need for computer-aided diagnosis techniques to aid in the streamlined care of patients and individualized treatment decisions [56]. To enable early treatment of acute stroke, early and reliable diagnosis is required, which heavily relies on imaging. The vast majority of strokes are of ischemic origin, caused by a blood clot occluding an artery resulting in oxygen deprivation of the brain tissue supplied by this artery. Typical causes are large vessel occlusion with or without thrombus dislodgement (e.g., carotid stenosis) or a cardiac cause resulting in embolies (e.g., atrial fibrillation). The less common subtype is hemorrhagic stroke, which has

substantially different etiology and is often caused by hypertension. Without early treatment of stroke, prognosis is poor. Each minute without treatment leads to loss of an estimated 1.8 million neurons [64]. Patients who enter the hospital with acute stroke symptoms often immediately undergo CT (or MR) scanning, even before detailed clinical evaluation of the patient [64]. Imaging here has three roles in decision making for treatment: (1) rule out hemorrhagic stroke, (2) establish the exact cause and the extent of ischemic stroke, and (3) determine a patient's suitability for (intra-arterial) treatment [33, 80]. Applications of machine learning for treatment decisions in stroke include identification of hemorrhage and early identification of imaging findings to determine the cause and extent of stroke and estimation of the time of onset. Time of onset is relevant since most current treatments aim for rapid reperfusion of ischemic tissue, either using intravenous thrombolytic medications or using endovascular techniques to mechanically remove the obstruction to blood flow, which should be performed within 4.5 h of stroke onset [56].

1.2 Prediction

Prediction or prognosis aims to understand the future “condition” of the patient, which can then be used for considering and planning therapeutic or lifestyle interventions proactively [22] that may slow the disease process or may reduce the risk for event recurrence. In addition, it can be used for effective patient management, for managing the expectations of patients and their caregivers [82], as well as for patient selection in clinical trials [35, 102]. We distinguish two main categories of prediction targets here: the natural disease course and patient outcomes after treatment.

Box 2: Prediction

Categories of prediction targets for which computer-aided models have or could have added value, with example brain disorders for which this is relevant as discussed in this chapter:

- **Natural disease course** Dementia and MS
- **Patient outcomes after treatment** MS, brain cancer, and stroke

Predicting the natural disease course, i.e., the future progression of the disease and its symptoms in a subject, is clinically relevant as it can aid care planning and managing the expectations of patients and caregivers about their future quality of life, physical health, and dependency [81]. Additionally, in disorders where treatment options are limited, it would improve future clinical trials

for new medication through identification of patients most likely to benefit from an effective treatment, i.e., those at early stages of disease who are likely to progress over the short-to-medium term (1–5 years) [83].

In dementia, prediction is challenging because of disease heterogeneity, i.e., differences in symptoms between patients along the disease process. For example, a patient can have either typical AD with memory problems or atypical AD with either language problems [43] or behavioral problems [99]. Moreover, patients with comparable brain atrophy may decline differently as the disease progresses, reflecting cognitive resilience due to genetic or lifestyle factors that may help to compensate for the level of atrophy [147]. Lastly, a similar symptom in two patients could be resulting from different diseases altogether. For example, a patient with mild cognitive impairment (MCI) either may have early stage dementia or may have cognitive impairment due to a different cause such as older age, injury, or a virus such as SARS-CoV-2 [44]. The latter, i.e., cognitive impairment due to non-degenerative disorders, is almost twice as prevalent as cognitive impairment due to dementia [106]. Here it is of interest to predict how the symptoms will develop over time for an individual; while patients without dementia may remain stable over time or even improve, the symptoms of patients with dementia typically worsen with time. Hence, the applications of machine learning in predicting the future course of dementia include the following: (i) predicting if a patient with cognitive impairment patient will develop dementia [138], (ii) predicting when the patient will reach a clinical dementia stage (i.e., duration of the prodromal disease phase) [83], and (iii) predicting the progression of biomarkers such as cognition and MRI measurements [61, 66].

In MS, especially in the early stages when patients experience clinical symptoms sporadically, prediction of the future disease course is highly relevant for care planning and expectation management. The early stage of MS, known as the relapsing-remitting phase, is characterized by sporadic inflammatory attacks on the neuronal protective coating called myelin. Over time, the recovery from these relapses becomes incomplete, resulting in permanent and progressive disability [144]. Because of this progressive nature and the variation between individuals, predicting the number of relapses and the time to permanent disability in a specific patient is highly important for care and treatment planning [18].

Next to prediction of the natural disease course, prediction of the future disease course after an intervention, i.e., **outcome prediction after treatment**, could be instrumental for planning of treatment and subsequent follow-up. This is of particular interest in MS where multiple treatment options are available. There are

currently 21 FDA-approved disease-modifying drugs available [27] that inhibit different aspects of pathological progression of MS mainly by immune modulation and sometimes through neuroprotection or remyelination. It is hence clinically highly relevant to choose the treatment option that an individual patient is expected to have most benefit from and to determine whether risks of second-line treatment are justified [131]. The same holds for stroke in the post-acute phase, where prediction of patient outcomes after treatment based on imaging may play a role for choosing between available treatments such as medication and rehabilitation therapy [80]. Here the focus is on the long term: reducing risk of recurrence and optimization of functioning. Computer-aided approaches can thus help in personalizing the treatment for a patient.

Predicting the outcomes after treatment is also of major interest for patients with brain tumors, and specifically in case of glioma where treatment response varies greatly across patients. Treatment usually consists of surgical resection followed by radiotherapy and/or chemotherapy. Almost invariably tumor recurrence or regrowth occurs; however, the question is when. In case of high-grade glioma (i.e., glioblastoma), tumor regrowth typically happens within a few months. In low-grade glioma, progression after treatment is often slower, and it may take years before any significant regrowth is detected; at some point, however, malignant transformation (to a high-grade glioma) may occur, leading to accelerated regrowth. As discussed in Subheading 1.1, computer-aided diagnosis methods can be used to identify the current tumor's genetic and histological profile, which already provides important prognostic information. Beyond this example of computer-aided differential diagnosis, machine learning methods can contribute in different ways by directly predicting outcomes after treatment [65, 127]. First, machine learning methods have shown promise to aid the differentiation between tumor progression and treatment-related abnormalities (pseudoprogression, radiation necrosis) [54, 65, 73, 127, 143]. Second, machine learning can be used to predict local relapse locations after radiotherapy, thus highlighting locations that should be targeted with a higher radiation dose, leading to personalized radiotherapy planning [114]. Third, a machine learning approach can predict local response to stereotactic radiosurgery of brain metastases, based on radiomics analysis of pretreatment MRI, where the outcome of interest (local tumor progression) was defined in terms of maximum axial diameter growth as measured on a follow-up scan [94]. Fourth, machine learning methods have been proposed for prediction of progression-free and overall survival, which aids care planning and managing the expectations of patients about their future [60, 108, 122, 127].

2 Method Evaluation

2.1 State-of-the-Art

Methodology for

Diagnosis and Prediction

For **early diagnosis in dementia**, a large body of research has been published on classification of subjects into AD, mild cognitive impairment (MCI), and normal aging [36, 113, 145]. Overall, classification methods show high performance for classification of AD patients and cognitively normal controls with an area under the receiver operating characteristic curve (AUC) of 85–98%. Reported performances are somewhat lower for early diagnosis in patients with MCI, i.e., prediction of imminent conversion to AD (AUC: 62–82%). Dementia classification is usually based on clinical diagnosis as a reference standard for training and validation [87], but biological diagnosis based on assessment of amyloid pathology with PET imaging or CSF has been increasingly used over the last years [53, 129]. Structural T1-weighted (T1w) MRI to quantify neuronal loss is the most commonly used biomarker, whereas the support vector machine (SVM) is the most commonly used classifier. For T1w, both voxel-based maps (e.g., voxel-based morphometry maps quantifying local gray matter density [62]) and region-based features [78] have been frequently used. While using only region-based volumes may limit performance, combining those with regional shape and texture has been shown to perform competitively with using voxel-wise maps [13, 15, 24]. Using multimodal imaging such as FDG-PET or DTI in addition to structural MRI may have added value over structural MRI only, but limited data is available [76, 150]. Following the trends and successes in medical image analysis and machine learning, neural network classifiers –convolutional neural networks (CNN) in particular–have increasingly been used since a few years [16, 145], but have not been shown to significantly outperform conventional classifiers. In addition, data-driven disease progression models are being developed [101], which do not rely on a priori defined labels but instead derive disease progression in a data-driven way.

Regarding **differential diagnosis in dementia**, studies focus mostly on discriminating AD from other types of dementia. Differential diagnosis based on CSF and PET biomarkers of AD pathology has shown good performance for distinguishing AD from FTLD with sensitivities of 0.83 (p-tau/amylod- β ratio from CSF) and 0.87 (amyloid PET) [48, 111, 117]. In addition, machine learning approaches have been published based on either structural or multimodal MRI as region-wise or voxel-wise imaging features and generally SVM as a classifier, similar to those used for early diagnosis in dementia. These methods focused mostly on differential diagnosis of AD and FTLD and reported performances in the range of AUC = 0.75 – 0.85 [12, 14, 92, 110]. A few studies addressed differential diagnosis of AD and vascular

dementia (VaD) [151] or multiclass differential diagnosis (5+ classes including AD, FTLD, VaD, dementia with Lewy bodies, and subjective cognitive decline) [93, 133].

For **differential diagnosis in brain cancer**, numerous MRI-based machine learning approaches have been presented. These developments have partly been facilitated by the availability of several valuable public datasets; see, for example, the overviews in [89, 135]. Most literature is dedicated to glioma characterization, which is therefore discussed in more detail here. Studies vary in the choice of input MRI sequences (T1w pre- and post-contrast, FLAIR, T2w, diffusion-weighted imaging, perfusion-weighted imaging, MR spectroscopy, APT CEST), the machine learning methodology (ranging from conventional radiomics approaches with hand-crafted features derived from manual tumor segmentations to deep learning approaches that automatically segment the tumor), the classification target(s) (e.g., grade, IDH, 1p19q, and/or MGMT status), the selection of glioma subtypes on which the method is validated (e.g., only low-grade glioma, only high-grade glioma, or both), and the extent of validation performed (single train-test split, repeated cross-validation, internal versus external validation). A systematic review on the use of machine learning in neuro-oncology found four articles on glioma grading, and four articles on identifying genetic/molecular characteristics of glioma based on MRI [122]. Among those, only one study used convolutional neural networks as a machine learning tool—to predict 1p19q status in low-grade glioma [1]. A more recent systematic review identified 27 studies on glioma grading of which 6 used deep learning, and 48 studies on MRI-based estimation of genetic/molecular characteristics of which 8 used deep learning [19]. Another recent review dedicated to machine learning approaches for MRI-based glioma characterization found 12 studies on glioma grading of which 2 used deep learning, and 43 studies on molecular characterization out of which 10 used deep learning [41]. These numbers indicate a trend toward deep learning approaches as we see in the entire field, but with conventional machine learning approaches with pre-defined radiomics features still being used frequently. Regarding the performance, two recent systematic reviews performed a meta-analysis of studies on molecular characterization of glioma. Jian et al. [55] found a pooled sensitivity/specificity/AUC in the validation set of 0.85/0.83/0.90 for IDH status prediction (12 studies), and 0.70/0.72/0.75 for 1p19q status prediction (5 studies). For MGMT, sensitivities and specificities ranging from 0.70 to 0.88 were found in 3 studies reporting validation performance, not allowing a meta-analysis. Van Kempen et al. [136] reported a pooled AUC of 0.91 for IDH status prediction (7 studies), 0.75 for 1p19q status prediction (3 studies), and 0.87 for MGMT promoter status prediction (3 studies). Thus, while the studies applied somewhat different

criteria for inclusion in the meta-analysis and used different statistical analysis methods, they obtained similar performance estimates. Whereas both meta-analyses suggest promising accuracy for MRI-based MGMT promoter status prediction based on the results reported in literature, a comprehensive evaluation of deep learning approaches for MGMT promoter status prediction on the BraTS2021 dataset [6] yielded disappointing results, with AUCs ranging from 0.5 to 0.6 [120]. Also, the winning method of the BraTS2021 challenge achieved an AUC of 0.62 [8], suggesting that MGMT promoter status prediction from MRI is a very difficult task. Both systematic reviews [55, 136] also pointed out the low proportion of studies with external validation (10 out of 44 in [55] and 12 out of 60 in [136]). Figure 2, recreated based on [55], shows a number of other insightful statistics on the methodologies found in literature. Finally, both reviews also identified machine learning methods aimed at predicting other, less frequently considered molecular targets, including ATRX, TERT, EGFR, P53, and PTEN, indicating the broad range of possible future research directions in this area.

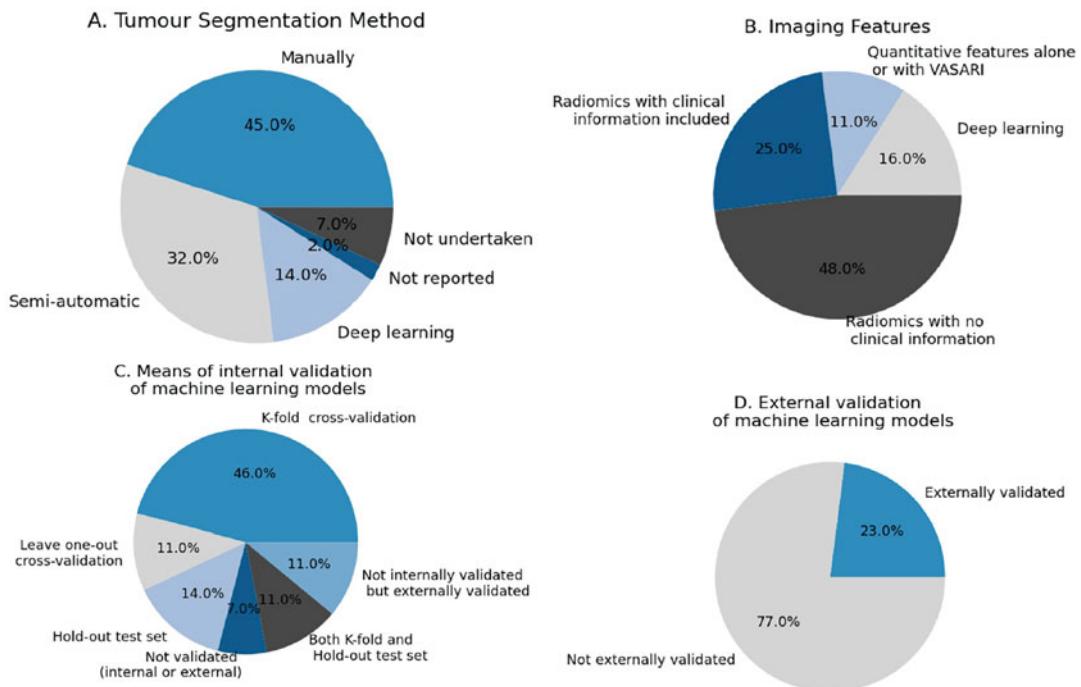


Fig. 2 Summary of tumor segmentation methods (a), types of imaging features (b), means of internal validation (c), and external validation (d) used by studies ($n=44$) investigating machine learning models for predicting genetic subtypes of glioma. VASARI, Visually Accessible Rembrandt Imaging. Recreated from [55]. Permission to reuse was kindly granted by the publishers

Beyond glioma characterization, other differential diagnosis problems in brain cancer are differentiation between glioma and lymphoma, between glioblastoma and metastasis, between different types of meningioma, and between glioma, meningioma, and pituitary tumors [19, 65, 122, 127, 149], with promising performances reported (AUC/accuracies around 90%). Of note, a recent study pointed out an important potential source of bias (the “Clever Hans effect”) in studies focused on differentiation between glioma, meningioma, and pituitary tumors, due to implicit radiologist input in the selection of the 2D slices in a commonly used benchmark dataset [142].

For decision making in stroke, different targets for machine learning based on imaging data have been identified, mostly focused at determining the cause and extent of stroke and to a lesser extent, on informing treatment decisions [56]. Regarding cause and extent of acute stroke, automatic lesion detection and identification of tissue-at-risk include the most important elements. These remain challenging as there is a lot of variation in lesion shape and location depending on time-from-symptom onset, vessel occlusion site, and collateral status [70]. Machine learning methods for segmentation and detection are increasingly successful (see Chapter 13). The step toward computer-aided diagnosis in stroke is also being taken using, for example, the CE-marked eASPECTS score [49], which is a machine learning-based assessment of the Alberta Stroke Program Early Computed Tomography Score (ASPECTS). This system for scoring acute ischemic damage to the brain has shown to be a simple, reliable, and strong predictor of functional outcome after stroke. Regarding treatment decisions, machine learning is used in several studies to determine whether a patient qualifies for a specific stroke treatment. For thrombolytic treatment, this qualification depends on time elapsed after symptom onset and treatment should be performed within 4.5 h. For this application, methods are developed that provide a binary estimation of stroke onset time (i.e., more or less than 4.5 h) based on either DWI and FLAIR [71] or perfusion-weighted imaging (CT or MR) [50]. Both approaches used a radiomics-like approach of feature extraction (e.g., intensity/gradient/texture based or using an autoencoder) followed by a machine learning classifier (support vector machine, random forest, and logistic regression). These machine learning methods had greater sensitivity than human readers using the standard procedure of DWI-FLAIR mismatch and comparable specificity. In addition, thrombolysis may cause the rare complication of symptomatic intracranial hemorrhage. Several machine learning methods have been developed to predict the risk of this complication achieving promising predictive performance, for example, using a support vector machine classifiers based on CT data (AUC = 0.74) [9].

For prediction of the future course of subjects at-risk of developing dementia, there are three frequently used approaches for defining the prediction problem at hand. First, predicting whether the patient will develop dementia. In specific diseases, measurement of proteins causing underlying pathology has shown to be very promising to identify patients in a prodromal disease state. Here, prediction is performed either using univariate analysis or using logistic regression with few variables as input. Blood-based P-Tau biomarker can predict incident AD within 4 years with an AUC of 0.78–0.83 [103], and CSF biomarkers and PET images of amyloid β and Tau can predict clinical progression of subjects in their prodromal AD state with an AUC of 0.94–0.96 [46]. Alternatively, in the absence of pathology-specific markers, MRI and cognitive markers of a patient together with machine learning approaches have been used to predict AD with an AUC of 0.70–0.83 [16, 23, 75, 141]. For a systematic review of the different machine learning methods developed for the purpose of predicting AD, see [5]. Support vector machines (SVM) and logistic regressions are the most used algorithms in the last decade (Fig. 3). In FTD, where it is currently not possible to measure the pathological proteins in body fluids, prediction based on a combination of biomarkers that are nonspecific to the underlying pathology is promising. This is demonstrated, for example, by van der Ende et al. [134], who predicted disease onset in familial FTD based on unspecific blood-based and CSF-based biomarkers using a disease progression model and identified presymptomatic subjects that developed dementia in the near future with an AUC of 0.85.

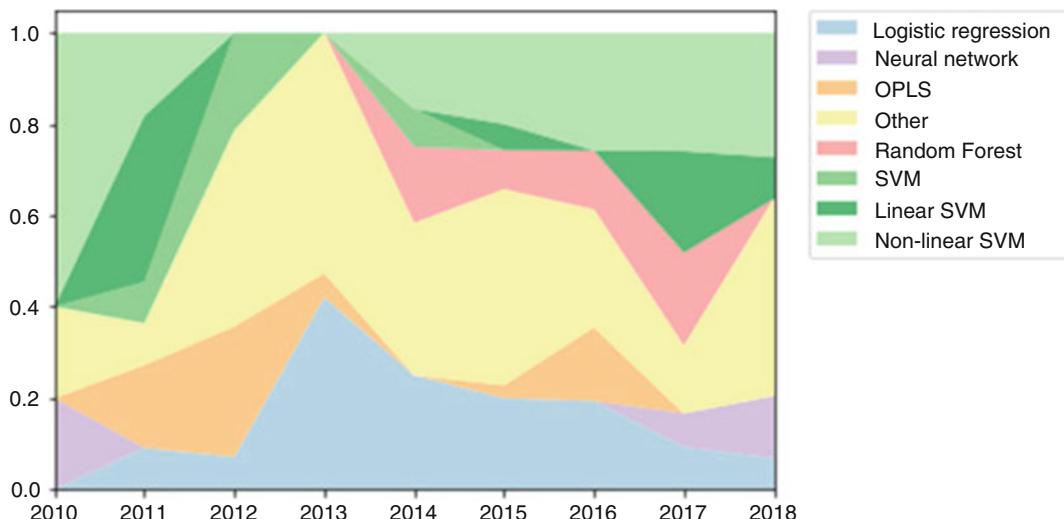


Fig. 3 Evolution with time of the use of various algorithms for predicting the progression of mild cognitive impairment. SVM with unknown kernel are simply noted as “SVM.” OPLS, orthogonal partial least square; SVM, support vector machine. Reproduced from [5]. Permission to reuse was kindly granted by the publishers

Second, predicting the time for conversion to dementia. While the previous problem predicts a dichotomous output variable, here it involves predicting a continuous variable of time to dementia. Bilgel et al. [11] predicted time to AD dementia with a mean error of <1.5 years. In the TADPOLE challenge, machine learning approaches to predict time for conversion to AD dementia of 33 participating teams have been assessed quantitatively [83]. Ansart et al. [5] strongly favor predicting the exact time for conversion to dementia and argue against predicting converters within a given time interval (e.g., within 3 years), because of the precision in the predictions. While this is indeed methodologically more elegant, the implications for clinical use and perception of patients regarding prediction precision and the inherent uncertainty remain to be established.

Third, prediction of disease markers could help to obtain insight into the clinical prognosis in an individual. Important disease markers are, for example, measures of global cognition (mini-mental state examination [MMSE] or Alzheimer's disease assessment scale [ADAS] scores), or salient imaging markers (volume of the brain ventricles or longitudinal Tau protein accumulation). ADAS scores could not be reliably predicted by any participating team in the TADPOLE challenge [83], but a recent disease progression model called *AD course map* [66] could predict ADAS scores (which is scored from 0 to 150) after 3 years with a mean absolute error of 7.6 points. *AD course map* could also predict MMSE scores (which is scored from 0 to 30) after 3 years with a mean absolute error of 3.2 points. While these predictions used MRI as input, Tau PET was recently shown to be more predictive of future MMSE scores using linear mixed models [100]. However, a thorough validation of this Tau PET-based prediction is lacking. Predicting salient imaging markers such as volume of the ventricles [83], volume of the hippocampus [66], or longitudinal Tau accumulation [72] is a promising topic. Identifying the most clinically useful target to be predicted, the imaging modality that has the best cost-benefit ratio for prognosis of a patient, and the method that best predicts it are all important questions that still need answers in the future.

Most **prediction methods in MS** focus on predicting either physical disability, cognitive impairment, or treatment response in imaging data of an individual patient [57]. Physical disability as measured by expanded disability status scale (EDSS, range 0–10) has been the most commonly used predictor variable as recently used in [104, 118]. An ensemble of classifiers consisting of convolutional neural networks, random forests, and manifold learning was reported to predict EDSS with a mean square error of 3.0 [118]. Cognitive impairment has been predicted either as a global measure of cognition or as specific cognitive domains such as attention or working memory [32]. For predicting treatment

response in MS, Signori et al. [124] used meta-analysis to identify subject characteristics that have higher treatment effects. In [34], the authors used an unsupervised disease progression model to identify subtypes of progression pathways in MS and found in post hoc analysis that one of the subtypes predicted better treatment effects. Current challenges in this evolving field of predicting treatment response in MS and future directions have been summarized in [37].

For the **prediction of patient outcomes after treatment of brain cancer**, most machine learning studies have focused on MRI-based prediction of progression-free survival or overall survival, which will therefore be discussed in more detail here. A systematic review by Sarkiss et al. identified nine articles on survival prediction in glioma, and two on survival prediction for patients after stereotactic radiosurgery of brain metastases [122]. A more recent systematic review by Buchlak et al. identified 17 studies on survival prediction with performance estimates (AUC or accuracy) mostly in the range 0.7–0.8 [19]. Among those, only one study reported results of external validation, predicting overall survival of patients with low-grade glioma, and obtained an AUC of 0.71 with a model combining radiomics with non-imaging features including age, resection extent, grade, and IDH status [21]. Random (survival) forests and support vector machines were most often used methods. One study used a CNN as a pre-trained feature extractor [69]. Other recent approaches using CNNs to extract features that are subsequently combined with other factors into a final prognostic model include [45, 51, 96]. The 2017/2018 editions of the well-known BraTS challenges also included a task on overall survival prediction, with best teams obtaining accuracies around 0.6 in a three-class classification setting distinguishing short-, mid-, and long-survivors, [7]. Here, it was also pointed out that conventional machine learning methods outperformed deep learning methods, likely due to the limited size of available datasets for training.

Beyond MRI-based methods, methods using histopathology images and/or genomics data as input for the machine learning model are also considered in the literature on outcome prediction for glioma patients. In one of the pioneering studies on digital pathology images of glioma, better prognostication was obtained with deep learning when pathology images were combined with genetic markers (IDH, 1p19q) [91]. Preliminary work on so-called radiopathomics in glioma is also available, supporting the notion that combining histology and radiology features improves prognostication (overall survival prediction) in glioma patients [115, 116].

2.2 Benchmarks and Challenges

For 15 years, grand challenges have been organized in the biomedical image analysis research field. These are international benchmarks in competition form that have the goal of objectively comparing algorithms for a specific task on the same clinically

representative data using the same evaluation protocol. In such challenges, the organizers supply reference data and evaluation measures on which researchers can evaluate their algorithms. Over the past years, the number and the impact of such grand challenges have increased [79]. Also in the field of computer-aided diagnosis and prediction, such grand challenges have been organized. For example, in the dementia field, four challenges have been organized focusing on early diagnosis [3, 15, 121] and predicting the natural disease course [3, 83, 121]. In general, algorithms winning the challenges performed rigorous data pre-processing and combined a wide range of input features [17]. In the field of brain cancer, the series of BraTS challenges has had a major impact [6, 7]. These benchmarks are instrumental to gaining insight into successful approaches and their potential for use in clinical practice and clinical trials.

2.3 Open-Source Software

Open-source machine learning software such as scikit-learn¹ and MONAI² have been fundamental to the development of this field of research. More specifically for computer-aided diagnosis and prediction in brain diseases, dedicated platforms are available such as Clinica [119], NeuroPredict [109], and PRoNTo [123]. We also see a trend of researchers publishing their scripts and trained classifiers with their publications in order to promote reproducibility.

3 Clinical Impact

There are multiple ways in which computer-aided diagnosis and prediction models can make an impact on clinical practice. Key areas of impact are in decision making for treatment and care, replacing invasive diagnostic procedures and patient selection for clinical trials. Here we will discuss to what extent these clinical needs are addressed by current methods.

First, the most direct impact is on decision making for treatment and care. This not only affects clinical care and treatment planning in patients with, for example, dementia, stroke, MS, or brain cancer but also is important for managing the expectations of patients and their caregivers. Although high performances are achieved for some related tasks such as dementia classification, validation of those results on external datasets and clinical cohorts is still very limited as well as knowledge on the robustness of the methods. For other applications, there is still room for performance improvement, and key factors in achieving that would be the combination of multimodal input and the availability of more well-

¹ <https://scikit-learn.org/>.

² <https://monai.io/>.

maintained and large-scale datasets for training and evaluation. In general, there is room for improvement in how well real clinical questions are addressed by current methodology. Second, machine learning models can have an impact by replacing invasive diagnostic procedures. This is especially relevant in brain cancer, where machine learning techniques based on imaging data are developed to predict, for example, genetic mutation status or tumor grade, thereby avoiding or reducing the need for biopsies [10, 152]. As a motivating example, MRI-based prediction of MGMT methylation status could be beneficial to guide treatment decisions. This is supported by findings from a population-based study assessing survival in 131 patients with radiological diagnosis of glioblastoma who did not undergo surgery and thus lacked (histological or molecular) tissue-based verification of the diagnosis [146]. While patients without treatment had extremely poor prognosis with median survival of 3.6 months, those who received upfront temozolamide treatment did significantly better (with median survival of 6.8 months). Since the response to temozolamide is known to be highly dependent on the MGMT status, MRI-based prediction of MGMT status could give insight into which patients would benefit from treatment avoiding the need for biopsies in patients too frail for tumor biopsy. Third, patient selection for clinical trials is relevant in diseases where no to limited options for treatment exist, such as dementia, or diseases where existing treatments are suboptimal for some patients, such as MS. This can boost the power of trials by enrolling, for example, individuals who are more likely to progress based on prediction models. Several pilot studies demonstrated the added value of machine learning models to select a subgroup of participants to increase sensitivity to the treatment using phase III trial data (e.g., for Alzheimer's disease treatment using donepezil or semagacestat) [35, 102]. This will ultimately reduce the size, duration, and cost of clinical trials.

The number of published methods is not evenly distributed over tasks. While many methods have been published on, for example, the classification of Alzheimer's disease patients versus controls, much fewer publications exist on differential diagnosis in dementia. In addition, there seems to be a mismatch in some applications between published classification methods and clinical needs, e.g., the clinically relevant problem of early diagnosis does not directly translate to the frequently studied classification task of established Alzheimer's disease versus healthy controls, but would instead require separation of early disease stage Alzheimer's disease patients from those that have cognitive complaints but not dementia.

Several approved machine learning products to assist diagnosis and prediction are making their way into clinical practice, in particular in the imaging domain. Van Leeuwen et al. evaluated 100 commercially available products for AI in radiology, of which 38 are related to brain diseases [137]. These include mostly segmentation,

quantification, and normative comparison for neurodegenerative diseases and detection of lesions for stroke and oncology. Most methods generate a sample radiologist report which can be inspected and modified. In dementia, for example, 17 reporting tools that use automated brain MRI segmentation software and normative reference data for single-subject comparison are regulatory approved for use in the memory clinic [105].

One of these is Quantib ND (Quantib BV, Rotterdam, the Netherlands),³ which is an approved commercial software that performs automatic segmentation into 20 brain regions as well as normative volumetry reference curves based on data of 5000 subjects from a population-based cohort. While Quantib ND and most other available tools use machine learning for brain segmentation, their output is not a diagnostic label produced by a machine learning algorithm. Another approved software, cDSI (Combinostics, Tampere, Finland),⁴ does output diagnostic labels as confidence scores in addition to segmentation and normative volumetry based on MRI. It uses univariate machine learning to normalize individual biomarkers of different modalities based on reference values of patient and control groups, color-codes these biomarkers to improve visualization of large-data datasets, and combines confidence scores based on individual biomarkers into one score [84, 85]. While cDSI is a machine learning tool for computer-aided diagnosis and prognosis, it does not exploit the power of machine learning to detect complex patterns in high-dimensional data but rather focuses on visualization and interpretability. Diagnosis and prediction algorithms that map high-dimensional input, i.e., images and other clinical data, to an outcome measure using machine learning have not yet made their way into clinical practice.

4 Roadmap for Clinical Translation

There are numerous challenges for clinical translation of computer-aided diagnosis and prediction methods. Some key items that should be on the roadmap for translation relate to large and standardized datasets, to technical and clinical validation, to interpretability by clinicians and patients, and to practical issues related to implementation. In this section, we will discuss these requirements and related developments and initiatives.

The first requirement for translation is **large and standardized datasets**. For a few brain disorders, one or multiple large datasets (i.e., up to 2500 participants) are available to train machine learning algorithms for diagnosis and prediction tasks, facilitated

³ quantib.com/solutions/quantib-nd.

⁴ combinostics.com/cdsi.

by large multicenter initiatives such as the Alzheimer’s Disease Neuroimaging Initiative (ADNI) or the Parkinson’s Progression Markers Initiative (PPMI). For validation in other cohorts and for development of algorithms in other diseases, there is only limited data available and a need for more (well-annotated) data exists. In particular, there is a need for validation data that reflect the reality of clinical routine with no to limited data harmonization and large variation in imaging protocols and data quality. Setting up such large-scale datasets is complex due to various reasons including obstacles in inter-institutional data sharing and a lack of funding for collection, curation, and labeling of data. To overcome these challenges, developments in research software and infrastructure may provide a solution by sharing easily reproducible algorithms rather than the data. Wrapping an algorithm in a container (e.g., Docker,⁵ Singularity [67]) and applying the algorithms locally to the data (at one site or multiple sites in a federated approach) enables method validation on large sets of data within the confines of the local institute’s firewalls. Such an approach could be also used for enabling training on larger datasets (i.e., federated learning [125]). Standardization of the data is important for eventual translation as it enables researchers to combine multiple datasets for development and validation of machine learning methods for diagnosis and prediction. Such standardization entails both data collection (e.g., diagnostic criteria, protocols for image acquisition, and clinical tests) and data organization (e.g., through open-source standards and platforms for data storage such as the Brain Imaging Data Structure (BIDS) and the Extensible Imaging Archive Toolkit (XNAT)).

Second, **technical and clinical validation** is a key focus area on the roadmap for translation. In the field of radiology, the quantitative neuroradiology initiative (QNI) framework has been developed as a model framework for translation defining the technical and clinical validation necessary to embed automated software into the clinical workflow [39]. Based on this framework, [105] reviewed the published evidence regarding commercial automated volumetric MRI tools for dementia diagnosis. For the 17 products identified, 11 companies have published some form of technical validation on their methods, but only 4 have published clinical validation in a dementia population. They concluded that there is a significant evidence gap in the literature regarding clinical validation and in-use evaluation. Whereas this review only addressed image volumetry in dementia, these findings likely extend to other brain diseases, applications, and modalities. Hence, there is a need for both retrospective and prospective studies validating algorithms in a clinical setting. In addition, performance metrics

⁵ www.docker.com.

used in validation studies should aim to capture real clinical applicability and address different aspects of the reliability of an algorithm, including accuracy, uncertainty estimation, reproducibility, and generalizability to other data. Standards for validation and reporting are provided by guidelines such as STARD-AI [130] and TRIPOD-AI.⁶

A third key item for clinical translation is **interpretability** by end users such as clinicians and patients. As clinicians have responsibility for the decisions related to care and treatment, they should have trust in a computer-aided diagnosis or prediction system and understand its outputs to an extent that they can rely on them for decision making and explanation to a patient. Performance metrics should aim to capture real clinical applicability and be understandable to intended users [59]. High validation performance is important for building trust in methods, but not sufficient by itself, since performance may reduce in individual cases because of unaccounted inter-individual such as comorbidities or population differences such as MRI scan protocol. Therefore, apart from model accuracy, relevant questions for interpretation are, for example, Is the model suitable for the data of this patient? What features contribute to the machine learning decision for this patient? How certain is the decision for this patient and can the algorithm know when it is uncertain about an individual's decision? Such questions are important and methods should be designed and implemented in a way that facilitates answers to such questions. This could be obtained by using interpretability methods on top of "black box" machine learning models or directly by using interpretable models. For the first category, many methods have been developed based on model weight visualization, feature map visualization, back-propagation methods, or perturbation of inputs (*see also Chapter 22*). For interpretable models, an example in the field of computer-aided diagnosis and prognosis is disease progression models [140, 148]. These data-driven models are designed specifically for neurodegenerative diseases and explain their decisions based on their estimate of the natural progression of the disease in the cohort (*see also Chapter 17*).

As a final key item, we will discuss **implementation feasibility**. For machine learning models to be actually used in practice, it is essential that models and reporting are integrated into the clinical workflow and that the sending and processing of clinical data and receiving results is fully automated. Current commercial products for automatic volumetry in dementia all reported to have implemented an integration with radiology systems and the clinical workflow. While validation of the workflows is limited [105], this does support the feasibility for machine learning in clinical practice.

⁶[osf.io/zyacb.](https://osf.io/zyacb/)

While these products integrate with the radiological workflow, a key challenge for the clinical translation of algorithms that use non-imaging clinical data (such as cognitive scores) as input is to also integrate with the clinical workflow of multidisciplinary diagnosis.

5 Final Summary and Conclusion

Computer-aided diagnosis and prediction of brain disorders is an important research area, with a wide variety of applications. While typically for these applications generic machine learning methods are used, domain knowledge of these brain disorders is crucial for selecting novel clinically relevant applications as well as for making domain-specific methodological improvements. Regarding diagnosis, clinical challenges are in early diagnosis of dementia and MS, differential diagnosis of dementia and brain cancer, and decision making for treatment in stroke. Regarding prediction, challenges are in the prediction of the natural disease course in dementia and MS, and the prediction of patient outcomes after treatment in stroke, brain cancer, and MS. Even though the disorders on which we focused are important avenues for impact, computer-aided diagnosis and prognosis would also be extremely useful in other disorders such as movement disorders for predicting response to treatment and side effects, epilepsy for predicting response to epilepsy surgery, and psychiatric disorders where diagnosis can be particularly difficult.

Key areas of impact are in (1) decision making for treatment and care in patients with dementia, stroke, MS, or brain cancer, (2) replacing invasive diagnostic procedures in brain cancer, and (3) patient selection for clinical trials in dementia and MS. While the first AI methods are making their way to clinical practice, diagnosis and prediction algorithms that map high-dimensional input, i.e., images and other clinical data, to an outcome measure using machine learning are not yet clinically available. To enable translation, major items on the roadmap relate to the availability of large and standardized datasets and technical and clinical validation of the developed machine learning methods. In addition, other important aspects are interpretability of the results by clinicians and patients, optimization of the diagnostic or treatment workflow in the clinic, and other practical issues related to implementation.

With this chapter, we aimed to provide a comprehensive overview, bringing together the clinical context of representative use cases of diagnosis and prediction in brain disorders and their state-of-the-art computer-aided methods. Future research should focus on bridging the identified gaps between clinical needs and the solutions brought by machine learning, to further improve decision making, treatment, and care in brain diseases.

6 Conflict of Interest

Quantib BV is a spin-off company of Erasmus MC. W.J.N. is a cofounder, part-time Chief Scientific Officer, and stockholder of Quantib BV. S.R.V, D.B., M.S., S.K., and E.E.B. are affiliated to Erasmus MC but have no personal relationships with or financial interest in Quantib BV. F.B. is a consultant to Combinostics.

Acknowledgements

V. Venkatraghavan is supported by JPND-funded E-DADS project (ZonMW project #733051106). W.J. Niessen and E.E. Bron are supported by Medical Delta Diagnostics 3.0: Dementia and Stroke. E.E. Bron acknowledges support from the Netherlands CardioVascular Research Initiative (Heart-Brain Connection: CVON2012-06, CVON2018-28).

References

1. Akkus Z, Ali I, Sedlář J, Agrawal JP, Parney IF, Giannini C, Erickson BJ (2017) Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from MR images using machine intelligence. *J Digit Imaging* 30(4): 469–476. <https://doi.org/10.1007/s10278-017-9984-3>
2. Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 7(3): 270–279. <https://doi.org/10.1016/j.jalz.2011.03.008>
3. Allen GI, Amoroso N, Anghel C, Balagurusamy V, Bare CJ et al (2016) Crowd-sourced estimation of cognitive decline and resilience in Alzheimer's disease. *Alzheimers Dement* 12(6):645–653. <https://doi.org/10.1016/j.jalz.2016.02.006>
4. Alzheimer's Association Report (2020) Alzheimer's disease facts and figures. *Alzheimers Dement* 16(3):391–460. <https://doi.org/10.1002/alz.12068>
5. Ansart M, Epelbaum S, Bassignana G, Boône A, Bottani S, Cattai T, Couronné R, Faouzi J, Koval I, Louis M, Thibreau-Sutre E, Wen J, Wild A, Burgos N, Dormont D, Colliot O, Durrleman S (2021) Predicting the progression of mild cognitive impairment using machine learning: a systematic, quantitative and critical review. *Med Image Anal* 67: 101848. <https://doi.org/10.1016/j.media.2020.101848>
6. Baid U, Ghodasara S, Mohan S, Bilello M, Calabrese E et al (2021) The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification
7. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M et al (2019) Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge
8. Bakas S, Baid U, Calabrese E, Carr C, Colak E, Farahani K, Flanders AE, Kalpathy-Cramer J, Kitamura FC, Menze B, Mongan J, Prevedello L, Rudie J, Shinohara RT (2021) RSNA-MICCAI brain tumor radiogenomic classification. [Link to the Kaggle challenge](#)
9. Bentley P, Ganeshalingam J, Carlton Jones AL, Mahady K, Epton S, Rinne P, Sharma P, Halse O, Mehta A, Rueckert D (2014) Prediction of stroke thrombolysis outcome using CT brain machine learning. *NeuroImage Clin* 4:635–640. <https://doi.org/10.1016/j.nicl.2014.02.003>
10. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, Allison T, Arnaout O, Abbosh C, Dunn IF, Mak RH, Tamimi RM, Tempany CM, Swanton C, Hoffmann U, Schwartz LH, Gillies RJ, Huang RY, Aerts HJWL (2019) Artificial

- intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin 69(2):127–157. <https://doi.org/10.3322/caac.21552>
11. Bilgel M, Jedynak BM, Initiative ADN (2019) Predicting time to dementia using a quantitative template of disease progression. Alzheimers Dement Diagn Assess Dis Monit 11(1): 205–215. <https://doi.org/10.1016/j.adm.2019.01.005>
 12. Bouts MJRJ, Möller C, Hafkemeijer A, van Swieten JC, Dopper E, van der Flier WM, Vrenken H, Wink AM, Pijnenburg YAL, Scheltens P, Barkhof F, Schouten TM, de Vos F, Feis RA, van der Grond J, de Rooij M, Rombouts SARB (2018) Single subject classification of alzheimer's disease and behavioral variant frontotemporal dementia using anatomical, diffusion tensor, and resting-state functional magnetic resonance imaging. J Alzheimer's Dis 62(4): 1827–1839. <https://doi.org/10.3233/jad-170893>
 13. Bron EE, Steketee RM, Houston GC, Oliver RA, Achterberg HC, Loog M, van Swieten JC, Hammers A, Niessen WJ, Smits M, Klein S, for the Alzheimer's Disease Neuroimaging Initiative (2014) Diagnostic classification of arterial spin labeling and structural MRI in presenile early stage dementia. Hum Brain Mapp 35(9):4916–4931. <https://doi.org/10.1002/hbm.22522>
 14. Bron E, Smits M, Papma J, Steketee R, Meijboom R, De Groot M, van Swieten J, Niessen W, Klein S (2016) Multiparametric computer-aided differential diagnosis of Alzheimer's disease and frontotemporal dementia using structural and advanced MRI. Eur Radiol 27(8):1–11. <https://doi.org/10.1007/s00330-016-4691-x>
 15. Bron EE, Smits M, van der Flier WM, Vrenken H, Barkhof F et al (2015) Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. NeuroImage 111:562–579. <https://doi.org/10.1016/j.neuroimage.2015.01.048>
 16. Bron EE, Klein S, Papma JM, Jiskoot LC, Venkatraghavan V, Linders J, Aalten P, De Deyn PP, Biessels GJ, Claassen JA, Middelkoop HA, Smits M, Niessen WJ, van Swieten JC, van der Flier WM, Ramakers IH, van der Lugt A (2021) Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. NeuroImage Clin 31: 102712. <https://doi.org/10.1016/j.nicl.2021.102712>
 17. Bron EE, Klein S, Reinke A, Papma JM, Maier-Hein L, Alexander DC, Oxtoby NP (2021) Ten years of image analysis and machine learning competitions in dementia
 18. Brown FS, Glasmacher SA, Kearns PKA, MacDougall N, Hunt D, Connick P, Chandran S (2020) Systematic review of prediction models in relapsing remitting multiple sclerosis. PLoS One 15(5):1–13. <https://doi.org/10.1371/journal.pone.0233575>
 19. Buchlak QD, Esmaili N, Leveque JC, Bennett C, Farrokhi F, Piccardi M (2021) Machine learning applications to neuroimaging for glioma detection and classification: an artificial intelligence augmented systematic review. J Clin Neurosci Off J Neurosurg Soc Australas 89:177–198. <https://doi.org/10.1016/j.jocn.2021.04.043>
 20. Chen C, Ou X, Wang J, Guo W, Ma X (2019) Radiomics-based machine learning in differentiation between glioblastoma and metastatic brain tumors. Front Oncol 9:806. <https://doi.org/10.3389/fonc.2019.00806>
 21. Choi YS, Ahn SS, Chang JH, Kang SG, Kim EH, Kim SH, Jain R, Lee SK (2020) Machine learning and radiomic phenotyping of lower grade gliomas: improving survival prediction. Eur Radiol 30(7):3834–3842. <https://doi.org/10.1007/s00330-020-06737-5>
 22. Crous-Bou M, Minguillón C, Gramunt N, Molinuevo JL (2017) Alzheimer's disease prevention: from risk factors to early intervention. Alzheimers Res Therapy 9(1):71. <https://doi.org/10.1186/s13195-017-0297-z>
 23. Cui Y, Liu B, Luo S, Zhen X, Fan M, Liu T, Zhu W, Park M, Jiang T, Jin JS, the Alzheimer's Disease Neuroimaging Initiative (2011) Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. PLoS One 6(7):1–10. <https://doi.org/10.1371/journal.pone.0021896>
 24. Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert MO, Chupin M, Benali H, Colliot O (2011) Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the adni database. NeuroImage 56(2):766–781. <https://doi.org/10.1016/j.neuroimage.2010.06.013>. Multivariate Decoding and Brain Reading
 25. Dekker I, Schoonheim MM, Venkatraghavan V, Eijlers AJ, Brouwer I, Bron EE, Klein S, Wattjes MP, Wink AM, Geurts JJ, Uitdehaag BM, Oxtoby NP, Alexander DC, Vrenken H, Killestein J, Barkhof F,

- Wottschel V (2021) The sequence of structural, functional and cognitive changes in multiple sclerosis. *NeuroImage Clin* 29: 102550. <https://doi.org/10.1016/j.nicl.2020.102550>
26. Delfanti RL, Piccioni DE, Handwerker J, Bahrami N, Krishnan A, Karunamuni R, Hattangadi-Gluth JA, Seibert TM, Srikant A, Jones KA, Snyder VS, Dale AM, White NS, McDonald CR, Farid N (2017) Imaging correlates for the 2016 update on WHO classification of grade II/III gliomas: Implications for IDH, 1p/19q and ATRX status. *J Neurooncol* 135(3):601–609. <https://doi.org/10.1007/s11060-017-2613-7>
27. Disease Modifying Therapies (2021) Disease modifying therapies for MS. National Multiple Sclerosis Society, New York, pp 3–21
28. Dubbink HJ, Atmodimedjo PN, Kros JM, French PJ, Sanson M, Idbah A, Wesseling P, Enting R, Spliet W, Tijssen C, Dinjens WNM, Gorlia T, van den Bent MJ (2015) Molecular classification of anaplastic oligodendrogloma using next-generation sequencing: a report of the prospective randomized EORTC Brain Tumor Group 26951 phase III trial. *Neuro-Oncology* 18(3):388–400. <https://doi.org/10.1093/neuonc/nov182>
29. Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL et al (2014) Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *Lancet Neurol* 13(6):614–629. [https://doi.org/10.1016/s1474-4422\(14\)70090-0](https://doi.org/10.1016/s1474-4422(14)70090-0)
30. Ducharme S, Price BH, Larvie M, Dougherty DD, Dickerson BC (2015) Clinical approach to the differential diagnosis between behavioral variant frontotemporal dementia and primary psychiatric disorders. *Am J Psychiatry* 172(9):827–837. <https://doi.org/10.1176/appi.ajp.2015.14101248>
31. Eckel-Passow JE, Lachance DH, Molinaro AM, Walsh KM, Decker PA et al (2015) Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. *N Engl J Med* 372(26):2499–2508. <https://doi.org/10.1056/NEJMoa1407279>
32. Eijlers AJC, van Geest Q, Dekker I, Steenwijk MD, Meijer KA, Hulst HE, Barkhof F, Uitdehaag BMJ, Schoonheim MM, Geurts JJG (2018) Predicting cognitive decline in multiple sclerosis: a 5-year follow-up study. *Brain* 141(9):2605–2618. <https://doi.org/10.1093/brain/awy202>
33. El-Koussy M, Schroth G, Brekenfeld C, Arnold M (2014) Imaging of acute ischemic stroke. *Eur Neurol* 72(5–6):309–316. <https://doi.org/10.1159/000362719>
34. Eshaghi A, Young A, Wijeratne P, Prados F, Arnold D, Narayanan S, Guttmann C, Barkhof F, Alexander D, Thompson A, Chard D, Ciccarelli O (2021) Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. *Nat Commun* 12(1). <https://doi.org/10.1038/s41467-021-22265-2>
35. Ezzati A, Lipton RB, Alzheimer's Disease Neuroimaging Initiative (2020) Machine learning predictive models can improve efficacy of clinical trials for Alzheimer's disease. *J Alzheimers Dis* 74(1):55–63. <https://doi.org/10.3233/jad-190822>
36. Falahati F, Westman E, Simmons A (2014) Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. *J Alzheimers Dis* 41(3):685–708. <https://doi.org/10.3233/jad-131928>
37. Gasperini C, Prosperini L, Tintoré M, Soriano MP, Filippi M, Rio J, Palace J, Rocca MA, Ciccarelli O, Barkhof F, Sastre-Garriga J, Vrenken H, Frederiksen JL, Yousry TA, Enzinger C, Rovira A, Kappos L, Pozzilli C, Montalban X, De Stefano N, The MAGNIMS Study Group (2019) Unraveling treatment response in multiple sclerosis. *Neurology* 92(4):180–192. <https://doi.org/10.1212/WNL.0000000000006810>
38. Gessler F, Bernstock JD, Braczynski A, Lescher S, Baumgarten P, Harter PN, Mittelbronn M, Wu T, Seifert V, Senft C (2019) Surgery for glioblastoma in light of molecular markers: impact of resection and MGMT promoter methylation in newly diagnosed IDH-1 wild-type glioblastomas. *Neurosurgery* 84(1):190–197. <https://doi.org/10.1093/neuros/nyy049>
39. Goodkin O, Pemberton H, Vos SB, Prados F, Sudre CH, Moggridge J, Cardoso MJ, Ourselin S, Bisdas S, White M, Yousry T, Thornton J, Barkhof F (2019) The quantitative neuroradiology initiative framework: application to dementia. *Br J Radiol* 92(1101):20190365. <https://doi.org/10.1259/bjr.20190365>, pMID: 31368776
40. Gordon BA, Blazey TM, Su Y, Hari-Raj A, Dincer A et al (2018) Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer's disease: a longitudinal study. *Lancet Neurol* 17(3):241–250. [https://doi.org/10.1016/S1474-4422\(18\)30028-0](https://doi.org/10.1016/S1474-4422(18)30028-0)
41. Gore S, Chougule T, Jagtap J, Saini J, Ingalkar M (2021) A review of radiomics and deep predictive modeling in glioma characterization. *Acad Radiol* 28(11):1599–1621.

- <https://doi.org/10.1016/j.acra.2020.06.016>
42. Gorelick PB, Scuteri A, Black SE, Decarli C, Greenberg SM et al (2011) Vascular contributions to cognitive impairment and dementia: a statement for healthcare professionals from the american heart association/american stroke association. *Stroke* 42(9):2672–2713. <https://doi.org/10.1161/STR.0b013e3182299496>
43. Gorno-Tempini ML, Brambati SM, Ginex V, Ogar J, Dronkers NF, Marcone A, Perani D, Garibotto V, Cappa SF, Miller BL (2008) The logopenic/phonological variant of primary progressive aphasia. *Neurology* 71(16):1227–1234. <https://doi.org/10.1212/01.wnl.0000320506.79811.da>
44. Hampshire A, Trender W, Chamberlain SR, Jolly AE, Grant JE, Patrick F, Mazibuko N, Williams SC, Barnby JM, Hellyer P, Mehta MA (2021) Cognitive deficits in people who have recovered from covid-19. *EClinicalMedicine* 39:101044. <https://doi.org/10.1016/j.eclinm.2021.101044>
45. Han W, Qin L, Bay C, Chen X, Yu KH, Miskin N, Li A, Xu X, Young G (2020) Deep transfer learning and radiomics feature prediction of survival of patients with high-grade gliomas. *Am J Neuroradiol* 41(1):40–48. <https://doi.org/10.3174/ajnr.A6365>
46. Hansson O, Seibyl J, Stomrud E, Zetterberg H, Trojanowski JQ, Bittner T, Lifke V, Corradini V, Eichenlaub U, Batrla R, Buck K, Zink K, Rabe C, Blennow K, Shaw LM, for the Swedish BioFINDER study group, Initiative ADN (2018) CSF biomarkers of Alzheimer's disease concord with amyloid- β pet and predict clinical progression: a study of fully automated immunoassays in biofinder and adni cohorts. *Alzheimers Dement* 14(11):1470–1481. <https://doi.org/10.1016/j.jalz.2018.01.010>
47. Harris JM, Thompson JC, Gall C, Richardson AM, Neary D, du Plessis D, Pal P, Mann DM, Snowden JS, Jones M (2015) Do NIA-AA criteria distinguish alzheimer's disease from frontotemporal dementia? *Alzheimers Dement* 11(2):207–215. <https://doi.org/10.1016/j.jalz.2014.04.516>
48. Hellwig S, Frings L, Bormann T, Vach W, Buchert R, Meyer PT (2019) Amyloid imaging for differential diagnosis of dementia: incremental value compared to clinical diagnosis and [^{18}F]fdg pet. *Eur J Nucl Med Mol Imaging* 46(2):312–323. <https://doi.org/10.1007/s00259-018-4111-3>
49. Herweh C, Ringleb PA, Rauch G, Gerry S, Behrens L, Möhlenbruch M, Gottorf R, Richter D, Schieber S, Nagel S (2016) Performance of e-ASPECTS software in comparison to that of stroke physicians on assessing CT scans of acute ischemic stroke patients. *Int J Stroke* 11(4):438–445. <https://doi.org/10.1177/1747493016632244>
50. Ho KC, Speier W, El-Saden S, Arnold CW (2017) Classifying acute ischemic stroke onset time using deep imaging features. *AMIA Ann Symp Proc* 2017:892–901
51. Huang H, Zhang W, Fang Y, Hong J, Su S, Lai X (2021) Overall survival prediction for gliomas using a novel compound approach. *Front Oncol* 11. <https://doi.org/10.3389/fonc.2021.724191>
52. Jack CR, Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, Shaw LM, Vemuri P, Wiste HJ, Weigand SD, Lesnick TG, Pankratz VS, Donohue MC, Trojanowski JQ (2013) Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol* 12(2):207–216. [https://doi.org/10.1016/S1474-4422\(12\)70291-0](https://doi.org/10.1016/S1474-4422(12)70291-0)
53. Jack Jr CR, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, Holtzman DM, Jagust W, Jessen F, Karlawish J, Liu E, Molinuevo JL, Montine T, Phelps C, Rankin KP, Rowe CC, Scheltens P, Siemers E, Snyder HM, Sperling R, Contributors, Elliott C, Masliah E, Ryan L, Silverberg N (2018) NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement* 14(4):535–562. <https://doi.org/10.1016/j.jalz.2018.02.018>
54. Jang BS, Jeon SH, Kim IH, Kim IA (2018) Prediction of pseudoprogression versus progression using machine learning algorithm in glioblastoma. *Sci Rep* 8(1):12516. <https://doi.org/10.1038/s41598-018-31007-2>
55. Jian A, Jang K, Manuguerra M, Liu S, Magnussen J, Di Ieva A (2021) Machine learning for the prediction of molecular markers in glioma on magnetic resonance imaging: a systematic review and meta-analysis. *Neurosurgery* 89(1):31–44. <https://doi.org/10.1093/neuros/nyab103>
56. Kamal H, Lopez V, Sheth SA (2018) Machine learning in acute ischemic stroke neuroimaging. *Front Neurol* 9:945. <https://doi.org/10.3389/fneur.2018.00945>
57. Kanber B, Nachev P, Barkhof F, Calvi A, Cardoso J, Cortese R, Prados F, Sudre C, Tur C, Ourselin S, Ciccarelli O (2019) High-dimensional detection of imaging

- response to treatment in multiple sclerosis. *NPJ Digit Med* 2(1). <https://doi.org/10.1038/s41746-019-0127-8>
58. Kaufmann M, Kuhle J, Puhan MA, Kamm CP, Chan A, Salmen A, Kesselring J, Calabrese P, Gobbi C, Pot C, Steinemann N, Rodgers S, von Wyl V, Swiss Multiple Sclerosis Registry (SMSR) (2018) Factors associated with time from first-symptoms to diagnosis and treatment initiation of multiple sclerosis in switzerland. *Mult Scler J Exp Transl Clin* 4(4): 2055217318814562. <https://doi.org/10.1177/2055217318814562>
 59. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D (2019) Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 17(1):195. <https://doi.org/10.1186/s12916-019-1426-2>
 60. Kickingereder P, Neuberger U, Bonekamp D, Piechotta PL, Götz M et al (2018) Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. *Neuro-Oncology* 20(6):848–857. <https://doi.org/10.1093/neuonc/nox188>
 61. Kim J, Park Y, Park S, Jang H, Kim HJ, Na DL, Lee H, Seo SW (2021) Prediction of tau accumulation in prodromal Alzheimer's disease using an ensemble machine learning approach. *Sci Rep* 11(1):5706. <https://doi.org/10.1038/s41598-021-85165-x>
 62. Klöppel S, Stonnington CM, Barnes J, Chen F, Chu C, Good CD, Mader I, Mitchell LA, Patel AC, Roberts CC, Fox NC, Jack J, Clifford R, Ashburner J, Frackowiak RSJ (2008) Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method. *Brain* 131(11): 2969–2974. <https://doi.org/10.1093/brain/awn239>
 63. Klöppel S, Abdulkadir A, Jack CR, Koutsouleris N, Mourão-Miranda J, Vemuri P (2012) Diagnostic neuroimaging across diseases. *NeuroImage* 61(2):457–463. <https://doi.org/10.1016/j.neuroimage.2011.11.002>. Neuroimaging: Then, Now and the Future
 64. Knight-Greenfield A, Nario JJQ, Gupta A (2019) Causes of acute stroke: a patterned approach. *Radiol Clin North Am* 57(6): 1093–1108. <https://doi.org/10.1016/j.rcl.2019.07.007>
 65. Kocher M, Ruge MI, Galldiks N, Lohmann P (2020) Applications of radiomics and machine learning for radiotherapy of malignant brain tumors. *Strahlenther Onkol* 196(10):856–867. <https://doi.org/10.1007/s00066-020-01626-8>
 66. Koval I, Bône A, Louis M, Lartigue T, Bottani S, Marcoux A, Samper-González J, Burgos N, Charlier B, Bertrand A, Epelbaum S, Colliot O, Allassonnière S, Durreiman S (2021) AD course map charts Alzheimer's disease progression. *Sci Rep* 11(1): 8020. <https://doi.org/10.1038/s41598-021-87434-1>
 67. Kurtzer GM, Sochat V, Bauer MW (2017) Singularity: scientific containers for mobility of compute. *PLoS One* 12(5):1–20. <https://doi.org/10.1371/journal.pone.0177459>
 68. Kuruppu DK, Matthews BR (2013) Young-onset dementia. *Semin Neurol* 33(4): 365–385. <https://doi.org/10.1055/s-0033-1359320>
 69. Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, Zhai G (2017) A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci Rep* 7(1):10353. <https://doi.org/10.1038/s41598-017-10649-8>
 70. Lee EJ, Kim YH, Kim N, Kang DW (2017) Deep into the brain: artificial intelligence in stroke imaging. *J Stroke* 19(3):277–285. <https://doi.org/10.5853/jos.2017.02054>
 71. Lee H, Lee EJ, Ham S, Lee HB, Lee JS, Kwon SU, Kim JS, Kim N, Kang DW (2020) Machine learning approach to identify stroke within 4.5 hours. *Stroke* 51(3):860–866. <https://doi.org/10.1161/strokeaha.119.027611>
 72. Leuzy A, Smith R, Cullen NC, Strandberg O, Vogel JW, Binette AP, Borroni E, Janelidze S, Ohlsson T, Jögi J, Ossenkoppele R, Palmqvist S, Mattsson-Carlgren N, Klein G, Stomrud E, Hansson O (2021) Biomarker-based prediction of longitudinal tau positron emission tomography in Alzheimer disease. *JAMA Neurol* <https://doi.org/10.1001/jamaneurol.2021.4654>
 73. Li M, Ren X, Dong G, Wang J, Jiang H, Yang C, Zhao X, Zhu Q, Cui Y, Yu K, Lin S (2021) Distinguishing pseudoprogression from true early progression in isocitrate dehydrogenase wild-type glioblastoma by interrogating clinical, radiological, and molecular features. *Front Oncol* 11. <https://doi.org/10.3389/fonc.2021.627325>
 74. Liquori M, Marrosu M, Pugliatti M, Giuliani F, De Robertis F, Cocco E, Zimatore G, Livrea P, Trojano M (2000) Age at onset in multiple sclerosis. *Neurol Sci J Ital Neurolog Soc Ital Soc Clin Neurophysiol* 21(4 Suppl 2):S825–S829. <https://doi.org/10.1007/s100720070020>
 75. Lin W, Tong T, Gao Q, Guo D, Du X, Yang Y, Guo G, Xiao M, Du M, Qu X, TADNI (2018)

- Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. *Front Neurosci* 12:777. <https://doi.org/10.3389/fnins.2018.00777>
76. Liu M, Cheng D, Wang K, Wang Y, Alzheimer's Disease Neuroimaging Initiative (2018) Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis. *Neuroinformatics* 16(3–4):295–308. <https://doi.org/10.1007/s12021-018-9370-4>
 77. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, Hawkins C, Ng HK, Pfister SM, Reifenberger G, Sofiotti R, von Deimling A, Ellison DW (2021) The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology* 23(8): 1231–1251. <https://doi.org/10.1093/neuonc/noab106>
 78. Magnin B, Mesrob L, Kinkinéhun S, Péligrini-Issac M, Colliot O, Sarazin M, Dubois B, Lehéricy S, Benali H (2009) Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology* 51(2):73–83. <https://doi.org/10.1007/s00234-008-0463-x>
 79. Maier-Hein LL, Eisenmann MM, Reinke AA, Onogur SS, Stankovic MM et al (2018) Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat Commun* 9(1). <https://doi.org/10.1038/s41467-018-07619-7>
 80. Mair G, Wardlaw JM (2014) Imaging of acute stroke prior to treatment: current practice and evolving techniques. *Br J Radiol* 87(1040): 20140216. <https://doi.org/10.1259/bjr.20140216>
 81. Mank A, van Maurik IS, Bakker ED, van de Glind EM, van Der Flier W, Visser LN (2020) Identifying patient-relevant endpoints in the progression of Alzheimer's disease. *Alzheimers Dement* 16(S6):e040866. <https://doi.org/10.1002/alz.040866>
 82. Mank A, van Maurik IS, Bakker ED, van de Glind EMM, Jönsson L, Kramberger MG, Novak P, Diaz A, Gove D, Scheltens P, van der Flier WM, Visser LNC (2021) Identifying relevant outcomes in the progression of Alzheimer's disease; what do patients and care partners want to know about prognosis? *Alzheimers Dement Transl Res Clin Interv* 7(1): e12189. <https://doi.org/10.1002/trc2.12189>
 83. Marinescu RV, Oxtoby NP, Young AL, Bron EE, Toga AW et al (2021) The Alzheimer's disease prediction of longitudinal evolution (TADPOLE) challenge: results after 1 year follow-up. *Mach Learn Biomed Imaging* 1
 84. Mattila J, Koikkalainen J, Virkki A, van Gils M, Lötjönen J (2012) Design and application of a generic clinical decision support system for multiscale data. *IEEE Trans Biomed Eng* 59(1):234–240. <https://doi.org/10.1109/TBME.2011.2170986>
 85. Mattila J, Soininen H, Koikkalainen J, Rueckert D, Wolz R, Waldemar G, Lötjönen J (2012) Optimizing the diagnosis of early Alzheimer's disease in mild cognitive impairment subjects. *J Alzheimers Dis* 32(4): 969–979. <https://doi.org/10.3233/jad-2012-120934>
 86. McAvoy M, Prieto PC, Kaczmarzyk JR, Fernández IS, McNulty J, Smith T, Yu KH, Gormley WB, Arnaout O (2021) Classification of glioblastoma versus primary central nervous system lymphoma using convolutional neural networks. *Sci Rep* 11(1): 15219. <https://doi.org/10.1038/s41598-021-94733-0>
 87. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack Jr CR, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (2011) The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 7(3): 263–269. <https://doi.org/10.1016/j.jalz.2011.03.005>
 88. Mehta D, Jackson R, Paul G, Shi J, Sabbagh M (2017) Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010–2015. *Expert Opin Investig Drugs* 26(6):735–739. <https://doi.org/10.1080/13543784.2017.1323868>
 89. Menze B, Isensee F, Wiest R, Wiestler B, Maier-Hein K, Reyes M, Bakas S (2021) Analyzing magnetic resonance imaging data from glioma patients using deep learning. *Comput Med Imaging Graph* 88:101828. <https://doi.org/10.1016/j.compmedimag.2020.101828>
 90. Miller JR (2004) The importance of early diagnosis of multiple sclerosis. *J Manag Care Pharm* 10(3 Suppl B):S4–11
 91. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, Brat DJ, Cooper LAD (2018) Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc*

- Natl Acad Sci 115(13):E2970–E2979. <https://doi.org/10.1073/pnas.1717139115>
92. Möller C, Pijnenburg YAL, van der Flier WM, Versteeg A, Tijms B, de Munck JC, Hafkemeijer A, Rombouts SARB, van der Grond J, van Swieten J, Dopper E, Scheltens P, Barkhof F, Vrenken H, Wink AM (2016) Alzheimer disease and behavioral variant frontotemporal dementia: automatic classification based on cortical atrophy for single-subject diagnosis. Radiology 279(3): 838–848. <https://doi.org/10.1148/radiol.2015150220>
93. Morin A, Samper-Gonzalez J, Bertrand A, Ströer S, Dormont D, Mendes A, Coupé P, Ahidjan J, Lévy M, Samri D, Hampel H, Dubois B, Teichmann M, Epelbaum S, Colliot O (2020) Accuracy of MRI classification algorithms in a tertiary memory center clinical routine cohort. J Alzheimers Dis 74(4): 1157–1166. <https://doi.org/10.3233/jad-190594>
94. Mouraviev A, Detsky J, Sahgal A, Ruschin M, Lee YK, Karam I, Heyn C, Stanisz GJ, Martel AL (2020) Use of radiomics for the prediction of local control of brain metastases after stereotactic radiosurgery. Neuro-Oncology 22(6):797–805. <https://doi.org/10.1093/neuonc/noaa007>
95. Muurling M, de Boer C, Kozak R, Religa D, Koychev I, Verheij H, Nies VJM, Duynstam A, Sood M, Fröhlich H, Hannesdottir K, Erdemli G, Lucivero F, Lancaster C, Hinds C, Stravopoulos TG, Nikolopoulos S, Kompatziaris I, Manyakov NV, Owens AP, Narayan VA, Aarsland D, Visser PJ, RADAR-AD Consortium (2021) Remote monitoring technologies in Alzheimer's disease: design of the RADAR-AD study. Alzheimers Res Therapy 13(1):89. <https://doi.org/10.1186/s13195-021-00825-4>
96. Nie D, Lu J, Zhang H, Adeli E, Wang J, Yu Z, Liu L, Wang Q, Wu J, Shen D (2019) Multi-channel 3d deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. Sci Rep 9(1): 1103. <https://doi.org/10.1038/s41598-018-37387-9>
97. Office for National Statistics (2019) Cancer survival in England: adult, stage at diagnosis and childhood – patients followed up to 2018. Dandy Booksellers Ltd., London
98. Okuda DT, Siva A, Kantarci O, Inglesse M, Katz I et al (2014) Radiologically isolated syndrome: 5-year risk for an initial clinical event. PLoS One 9(3):1–9. <https://doi.org/10.1371/journal.pone.0090509>
99. Ossenkoppele R, Singleton EH, Groot C, Dijkstra AA, Eikelboom WS et al (2021) Research criteria for the behavioral variant of Alzheimer disease: a systematic review and meta-analysis. JAMA Neurol <https://doi.org/10.1001/jamaneurol.2021.4417>
100. Ossenkoppele R, Smith R, Mattsson-Carlgren N, Groot C, Leuzy A et al (2021) Accuracy of tau positron emission tomography as a prognostic marker in preclinical and prodromal Alzheimer disease: a head-to-head comparison against amyloid positron emission tomography and magnetic resonance imaging. JAMA Neurol 78(8):961–971. <https://doi.org/10.1001/jamaneurol.2021.1858>
101. Oxtoby NP, Alexander DC, EuroPOND consortium (2017) Imaging plus x: multimodal models of neurodegenerative disease. Curr Opin Neurol 30(4):371–379. <https://doi.org/10.1097/wco.0000000000000460>
102. Oxtoby NP, Shand C, Cash DM, Alexander DC, Barkhof F, for the Alzheimer's Disease Neuroimaging Initiative, the Alzheimer's Disease Cooperative Study (2021) Targeted screening for alzheimer's disease clinical trials using data-driven disease progression models. medRxiv. <https://doi.org/10.1101/2021.01.29.21250773>
103. Palmqvist S, Tideman P, Cullen N, Zetterberg H, Blennow K, Alzheimer's Disease Neuroimaging Initiative, Dage JL, Stomrud E, Janelidze S, Mattsson-Carlgren-N, Hansson O (2021) Prediction of future Alzheimer's disease dementia using plasma phospho-tau combined with other accessible measures. Nat Med 27(6):1034–1042. <https://doi.org/10.1038/s41591-021-01348-z>
104. Pellegrini F, Copetti M, Sormani MP, Bovis F, de Moor C, Debray TP, Kieseier BC (2020) Predicting disability progression in multiple sclerosis: insights from advanced statistical modeling. Mult Scler J 26(14):1828–1836. <https://doi.org/10.1177/1352458519887343> PMID: 31686590
105. Pemberton HG, Zaki LAM, Goodkin O, Das RK, Steketee RME, Barkhof F, Vernooyj MW (2021) Technical and clinical validation of commercial automated volumetric MRI tools for dementia diagnosis-a systematic review. Neuroradiology 63(11):1773–1789. <https://doi.org/10.1007/s00234-021-02746-3>
106. Plassman BL, Langa KM, McCammon RJ, Fisher GG, Potter GG, Burke JR, Steffens DC, Foster NL, Giordani B, Unverzagt FW, Welsh-Bohmer KA, Heeringa SG, Weir DR,

- Wallace RB (2011) Incidence of dementia and cognitive impairment, not dementia in the united states. *Ann Neurol* 70(3):418–426. <https://doi.org/10.1002/ana.22362>
107. Prince M, Bryce R, Ferri C (2011) World Alzheimer Report 2011: The benefits of early diagnosis and intervention. *Alzheimers Dis Int*
108. Qiu X, Gao J, Yang J, Hu J, Hu W, Kong L, Lu JJ (2020) A comparison study of machine learning (random survival forest) and classic statistic (cox proportional hazards) for predicting progression in high-grade glioma after proton and carbon ion radiotherapy. *Front Oncol* 10:2311. <https://doi.org/10.3389/fonc.2020.551420>
109. Raamana PR (2017) NeurOpredict: easy machine learning and standardized predictive analysis of biomarkers 1058993. <https://doi.org/10.5281/zenodo.1058993>
110. Raamana PR, Rosen H, Miller B, Weiner MW, Wang L, Beg MF (2014) Three-class differential diagnosis among alzheimer disease, frontotemporal dementia, and controls. *Front Neurol* 5:71. <https://doi.org/10.3389/fneur.2014.00071>
111. Rabinovici G, Rosen H, Alkalay A, Kornak J, Furst A, Agarwal N, Mormino E, O’Neil J, Janabi M, Karydas A, Growdon M, Jang J, Huang E, DeArmond S, Trojanowski J, Grinberg L, Gorno-Tempini M, Seeley W, Miller B, Jagust W (2011) Amyloid vs FDG-PET— in the differential diagnosis of AD and FTLD. *Neurology* 77(23): 2034–2042. <https://doi.org/10.1212/WNL.0b013e31823b9c5e>. ISSN 0028-3878
112. Rascovsky K, Hodges JR, Knopman D, Mendez MF, Kramer JH et al (2011) Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain* 134(9):2456–2477. <https://doi.org/10.1093/brain/awr179>
113. Rathore S, Habes M, Iftikhar MA, Shacklett A, Davatzikos C (2017) A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer’s disease and its prodromal stages. *NeuroImage* 155:530–548. <https://doi.org/10.1016/j.neuroimage.2017.03.057>
114. Rathore S, Akbari H, Doshi J, Shukla G, Rozycski M, Bilello M, Lustig RA, Davatzikos CA (2018) Radiomic signature of infiltration in peritumoral edema predicts subsequent recurrence in glioblastoma: implications for personalized radiotherapy planning. *J Med Imaging* 5(2):1–10. <https://doi.org/10.1117/1.JMI.5.2.021219>
115. Rathore S, Iftikhar MA, Gurcan MN, Mourelatos Z (2019) Radiopathomics: integration of radiographic and histologic characteristics for prognostication in glioblastoma
116. Rathore S, Nasrallah M, Mourelatos Z (2019) NIMG-76. radiopathomics: integration of radiographic and histologic characteristics for prognostication in glioblastoma. *Neuro-Oncology* 21(Supplement 6):vi178–vi179. <https://doi.org/10.1093/neuonc/noz175.745>
117. Rivero-Santana A, Ferreira D, Perestelo-Pérez L, Westman E, Wahlund LO, Sarriá A, Serrano-Aguilar P (2017) Cerebrospinal fluid biomarkers for the differential diagnosis between Alzheimer’s disease and frontotemporal lobar degeneration: systematic review, HSROC analysis, and confounding factors. *J Alzheimers Dis* 55(2):625–644. <https://doi.org/10.3233/Jad-160366>
118. Roca P, Attye A, Colas L, Tucholka A, Rubini P et al (2020) Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. *Diagn Interv Imaging* 101(12):795–802. <https://doi.org/10.1016/j.diii.2020.05.009>
119. Routier A, Burgos N, Díaz M, Bacci M, Bottani S, El-Rifai O, Fontanella S, Gori P, Guillón J, Guyot A, Hassanalí R, Jacquemont T, Lu P, Marcoux A, Moreau T, Samper-González J, Teichmann M, Thibaut-Sutre E, Vaillant G, Wen J, Wild A, Habert MO, Durrleman S, Colliot O (2021) Clinica: an open-source software platform for reproducible clinical neuroscience studies. *Front Neuroinform* 15:39. <https://doi.org/10.3389/fninf.2021.689675>
120. Saeed N, Hardan S, Abutalip K, Yaqub M (2022) Is it possible to predict mgmt promoter methylation from brain tumor MRI scans using deep learning models?
121. Sarica A, Cerasa A, Quattrone A, Calhoun V (2018) Editorial on special issue: machine learning on MCI. *J Neurosci Methods* 302: 1–2. <https://doi.org/10.1016/j.jneumeth.2018.03.011>. A machine learning neuroimaging challenge for automated diagnosis of Alzheimer’s disease
122. Sarkiss CA, Germano IM (2019) Machine learning in neuro-oncology: can data analysis from 5346 patients change decision-making paradigms? *World Neurosurg* 124:287–294. <https://doi.org/10.1016/j.wneu.2019.01.046>
123. Schrouff J, Rosa M, Rondina J, Marquand A, Chu C, Ashburner J, Phillips C, Richiardi J, Mourão-Miranda J (2013) Pronto: pattern recognition for neuroimaging toolbox. *Neuroinformatics* 11(3):319–337. <https://doi.org/10.1007/s12021-013-9178-1>

124. Signori A, Schiavetti I, Gallo F, Sormani MP (2015) Subgroups of multiple sclerosis patients with larger treatment benefits: a meta-analysis of randomized trials. *Eur J Neurol* 22(6):960–966. <https://doi.org/10.1111/ene.12690>
125. Silva S, Altmann A, Gutman B, Lorenzi M (2020) Fed-biomed: a general open-source frontend framework for federated learning in healthcare. In: Albarqouni S, Bakas S, Kamnitsas K, Cardoso MJ, Landman B, Li W, Milletari F, Rieke N, Roth H, Xu D, Xu Z (eds) Domain adaptation and representation transfer, and distributed and collaborative learning. Springer International Publishing, Cham, pp 201–210
126. Simblett S, Matcham F, Curtis H, Greer B, Polhemus A, Novák J, Ferrao J, Gamble P, Hotopf M, Narayan V, Wykes T, Remote Assessment of Disease and Relapse – Central Nervous System (RADAR-CNS) Consortium (2020) Patients' measurement priorities for remote measurement technologies to aid chronic health conditions: qualitative analysis. *JMIR Mhealth Uhealth* 8(6):e15086. <https://doi.org/10.2196/15086>
127. Singh G, Manjila S, Sakla N, True A, Wardeh AH, Beig N, Vaysberg A, Matthews J, Prasanna P, Spektor V (2021) Radiomics and radiogenomics in gliomas: a contemporary update. *Br J Cancer* 125(5):641–657. <https://doi.org/10.1038/s41416-021-01387-w>
128. Smits M (2016) Imaging of oligodendrogloma. *Br J Radiol* 89(1060):20150857. <https://doi.org/10.1259/bjr.20150857>
129. Son HJ, Oh JS, Oh M, Kim SJ, Lee JH, Roh JH, Kim JS (2020) The clinical feasibility of deep learning-based classification of amyloid pet images in visually equivocal cases. *Eur J Nucl Med Mol Imaging* 47(2):332–341. <https://doi.org/10.1007/s00259-019-04595-y>
130. Sounderahaj V, Ashrafian H, Golub RM, Shetty S, De Fauw J et al (2021) Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 11(6). <https://doi.org/10.1136/bmjopen-2020-047709>
131. Stühler E, Braune S, Lionetto F, Heer Y, Jules E, Westermann C, Bergmann A, van Hövell P, Group NS, et al (2020) Framework for personalized prediction of treatment response in relapsing remitting multiple sclerosis. *BMC Med Res Methodol* 20. <https://doi.org/10.1186/s12874-020-0906-6>
132. Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T et al (2018) Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria. *Lancet Neurol* 17(2):162–173. [https://doi.org/10.1016/S1474-4422\(17\)30470-2](https://doi.org/10.1016/S1474-4422(17)30470-2)
133. Tong T, Ledig C, Guerrero R, Schuh A, Koikkalainen J, Tolonen A, Rhodius H, Barkhof F, Tijms B, Lemstra AW, Soininen H, Remes AM, Waldemar G, Hasselbalch S, Mecocci P, Baroni M, Lötjönen J, van der Flier W, Rueckert D (2017) Five-class differential diagnostics of neurodegenerative diseases using random undersampling boosting. *NeuroImage Clin* 15:613–624. <https://doi.org/10.1016/j.nicl.2017.06.012>
134. van der Ende EL, Bron EE, Poos JM, Jiskoot LC, Panman JL et al (2021) A data-driven disease progression model of fluid biomarkers in genetic frontotemporal dementia. *Brain*. <https://doi.org/10.1093/brain/awab382>, awab382
135. van der Voort SR, Incekara F, Wijnenga MMJ, Kapsas G, Gahrmann R, Schouten JW, Tewarie RN, Lycklama GJ, Hamer PCDW, Eijgelaar RS, French PJ, Dubbink HJ, Vincent AJPE, Niessen WJ, van den Bent MJ, Smits M, Klein S (2020) Who 2016 subtyping and automated segmentation of glioma using multi-task deep learning
136. van Kempen EJ, Post M, Mannil M, Kusters B, ter Laan M, Meijer FJA, Henssen DJHA (2021) Accuracy of machine learning algorithms for the classification of molecular features of gliomas on MRI: a systematic literature review and meta-analysis. *Cancers* 13(11). <https://doi.org/10.3390/cancers13112606>
137. van Leeuwen KG, Schalekamp S, Rutten MJ, van Ginneken B, de Rooij M (2021) Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol* 31(6):3797–3804. <https://doi.org/10.1007/s00330-021-07892-z>
138. van Maurik IS, Vos SJ, Bos I, Bouwman FH, Teunissen CE et al (2019) Biomarker-based prognosis for people with mild cognitive impairment (ABIDE): a modelling study. *Lancet Neurol* 18(11):1034–1044. [https://doi.org/10.1016/S1474-4422\(19\)30283-2](https://doi.org/10.1016/S1474-4422(19)30283-2)
139. van Vliet D, de Vugt ME, Bakker C, Pijnenburg YAL, Vernooij-Dassen MJFJ, Koopmans RTCM, Verhey FRJ (2013) Time to diagnosis in young-onset dementia as compared with late-onset dementia. *Psychol Med* 43(2):

- 423–432. <https://doi.org/10.1017/S0033291712001122>
140. Venkatraghavan V, Bron EE, Niessen WJ, Klein S (2019) Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. *NeuroImage* 186:518–532
141. Venkatraghavan V, Dubost F, Bron EE, Niessen WJ, de Bruyne M, Klein S (2019) Event-based modeling with high-dimensional imaging biomarkers for estimating spatial progression of dementia. In: Chung ACS, Gee JC, Yushkevich PA, Bao S (eds) *Information processing in medical imaging*. Springer International Publishing, Cham, pp 169–180
142. Wallis D, Buvat I (2022) Clever hans effect found in a widely used brain tumour MRI dataset. *Med Image Anal* 77:102368. <https://doi.org/10.1016/j.media.2022.102368>
143. Wang K, Qiao Z, Zhao X, Li X, Wang X, Wu T, Chen Z, Fan D, Chen Q, Ai L (2020) Individualized discrimination of tumor recurrence from radiation necrosis in glioma patients using an integrated radiomics-based model. *Eur J Nucl Med Mol Imaging* 47(6): 1400–1411. <https://doi.org/10.1007/s00259-019-04604-0>
144. Weinshenker BG (1994) Natural history of multiple sclerosis. *Ann Neurol* 36(S1):S6–S11. <https://doi.org/10.1002/ana.410360704>
145. Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Dormont D, Durrleman S, Burgos N, Colliot O (2020) Convolutional neural networks for classification of alzheimer's disease: overview and reproducible evaluation. *Med Image Anal* 63:101694. <https://doi.org/10.1016/j.media.2020.101694>
146. Werlenius K, Fekete B, Blomstrand M, Carén H, Jakola AS, Rydenhag B, Smits A (2020) Patterns of care and clinical outcome in assumed glioblastoma without tissue diagnosis: a population-based study of 131 consecutive patients. *PLoS One* 15(2):1–14. <https://doi.org/10.1371/journal.pone.0228480>
147. Yao T, Sweeney E, Nagorski J, Shulman JM, Allen GI (2020) Quantifying cognitive resilience in Alzheimer's disease: the Alzheimer's disease cognitive resilience score. *PLoS One* 15(11):1–21. <https://doi.org/10.1371/journal.pone.0241707>
148. Young AL, Oxtoby NP, Daga P, Cash DM, Fox NC, Ourselin S, Schott JM, Alexander DC (2014) A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* 137(9):2564–2577
149. Zegers C, Posch J, Traverso A, Eekers D, Postma A, Backes W, Dekker A, van Elmpt W (2021) Current applications of deep-learning in neuro-oncological MRI. *Ital Assoc Biomed Phys* 83:161–173. <https://doi.org/10.1016/j.ejmp.2021.03.003>
150. Zhang D, Wang Y, Zhou L, Yuan H, Shen D (2011) Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage* 55(3):856–867. <https://doi.org/10.1016/j.neuroimage.2011.01.008>
151. Zheng Y, Guo H, Zhang L, Wu J, Li Q, Lv F (2019) Machine learning-based framework for differential diagnosis between vascular dementia and Alzheimer's disease using structural MRI features. *Front Neurol* 10:1097. <https://doi.org/10.3389/fneur.2019.01097>
152. Zhou M, Scott J, Chaudhury B, Hall L, Goldgof D, Yeom K, Iv M, Ou Y, Kalpathy-Cramer J, Napel S, Gillies R, Gevaert O, Gatenby R (2018) Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches. *Am J Neuroradiol* 39(2): 208–216. <https://doi.org/10.3174/ajnr.A5391>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 16

Subtyping Brain Diseases from Imaging Data

Junhao Wen, Erdem Varol, Zhijian Yang, Gyujoon Hwang, Dominique Dwyer, Anahita Fathi Kazerooni, Paris Alexandros Lalousis, and Christos Davatzikos

Abstract

The imaging community has increasingly adopted machine learning (ML) methods to provide individualized imaging signatures related to disease diagnosis, prognosis, and response to treatment. Clinical neuroscience and cancer imaging have been two areas in which ML has offered particular promise. However, many neurologic and neuropsychiatric diseases, as well as cancer, are often heterogeneous in terms of their clinical manifestations, neuroanatomical patterns, or genetic underpinnings. Therefore, in such cases, seeking a single disease signature might be ineffectual in delivering individualized precision diagnostics. The current chapter focuses on ML methods, especially semi-supervised clustering, that seek disease subtypes using imaging data. Work from Alzheimer's disease and its prodromal stages, psychosis, depression, autism, and brain cancer are discussed. Our goal is to provide the readers with a broad overview in terms of methodology and clinical applications.

Key words Neuroimaging, Machine learning, Semi-supervised clustering, Heterogeneity

1 Introduction

There is a growing clinical evidence that structural and functional brain development and aging take heterogeneous paths within different subsets of the human population [1–3]. This heterogeneity has been relatively ignored in case-control study analyses, yielding a limited understanding of the diversity of underlying biological processes that might give rise to similar clinical phenotypes. The advent of high-throughput neuroimaging technologies and the concentrated efforts of the collection of large-scale datasets [4, 5] provide a unique opportunity to dissect the structural and functional heterogeneity of brain disorders in finer details and in an unbiased data-driven manner. A developing body of work that leverages ML and neuroimaging seeks disease subtypes of neuropsychiatric and neurodegenerative disorders, including Alzheimer's disease (AD) [6–11], schizophrenia [12, 13], and late-life depression [14].

Subtyping brain diseases is a clustering problem where the goal is to break down the set of patients into distinct and relatively homogeneous subgroups (i.e., subtypes). While this has been actively investigated in the computer science community, subtyping neuroimaging data is endowed with a unique set of obstacles, such as the “curse of dimensionality” and the confounding nuisance effects, such as global demographics and scanner differences. Furthermore, brain development and pathologies often progress along a continuum, e.g., from healthy state to preclinical stages to full-fledged disease [15], thereby modeling directly in the patient domain may lead to a biased clustering solution. Thus, to tackle these problems, some recent efforts have focused on developing semi-supervised [6, 8, 9, 16] and unsupervised clustering methods [10, 11]. Early studies mainly focused on unsupervised clustering methods, such as K-means [17] or hierarchical clustering [18], to derive data-driven subtypes using imaging data. However, such approaches directly partition the patients based on similarities/dissimilarities, potentially biased by confounding factors, such as demographics or heterogeneity caused by unrelated pathological processes. More recently, semi-supervised clustering methods [6, 8, 9, 16] have been proposed to tackle this problem from a novel angle. To seek a pathology-oriented clustering solution, semi-supervised approaches dissect disease heterogeneity by the “1-to- k ” mapping between the reference group (i.e., healthy control (CN)) and the subgroups of the patient group (i.e., the k subtypes). This approach presumably zooms into the heterogeneity of pathological processes rather than unwanted heterogeneity in general. Furthermore, confounding variations, such as demographics, are often ruled out in these approaches.

Aiming to provide the reader in the imaging and machine learning community with a broad guideline in terms of methodology and clinical applications, we organize the remainder of this chapter as follows. In Subheading 2, we provide a brief overview of clustering methods, including unsupervised and semi-supervised approaches. Subheading 3 discusses their applications in various neurological and neuropsychiatric disorders and diseases. Subheading 4 concludes the paper by discussing our main observations, methodological limitations, and future directions.

2 Methodological Development Using Machine Learning and Neuroimaging

Machine learning and neuroimaging have brought unprecedented opportunities to elucidate disease heterogeneity in various brain disorders and diseases [19]. Several trailblazing methodological papers have been recently published [9–11], challenging the conventional approach of patient stratification that puts all patients into the same bucket. Among these, unsupervised [10, 11] and semi-

supervised clustering methods [9] sought to derive biologically data-driven disease subtypes, but they anchor the modeling from distinct perspectives. For conciseness, let us note that our imaging dataset contains q healthy control (CN) samples $\mathbf{X}_r = [\mathbf{x}_1, \dots, \mathbf{x}_q], \mathbf{X}_r \in \mathbb{R}^{p \times q}$, representing our reference group, and n patient samples (PT) $\mathbf{X}_t = [\mathbf{x}_1, \dots, \mathbf{x}_m], \mathbf{X}_t \in \mathbb{R}^{p \times m}$, representing the target subtype population. We denote the whole population as a matrix \mathbf{X} that is organized by arranging each image as a vector per column $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_{q+m}], \mathbf{X} \in \mathbb{R}^{p \times (q+m)}$, where p is the number of features per image. We use binary labels to distinguish the patient and control groups, where 1 represents PT and -1 means CN. Disease subtyping sought to find the number of clusters (k) in the patient group that are neuroanatomically distinct while clinically relevant.

2.1 Unsupervised Clustering

Many recent efforts to discover the heterogeneous nature of brain diseases have investigated different unsupervised clustering algorithms [10, 11, 20–32]. Among these approaches, the key clustering methods are often K-means, hierarchical clustering, and nonnegative matrix factorization (NMF) (Fig. 1). In this

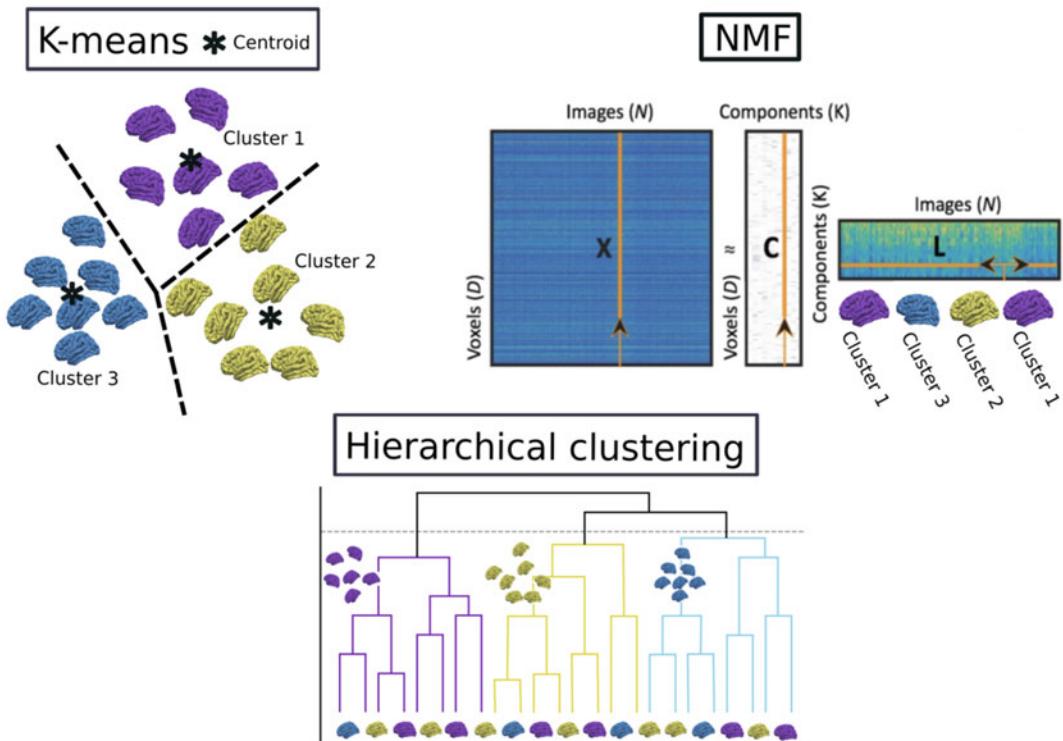


Fig. 1 Schematic diagram of representative unsupervised clustering methods, K-means, NMF, and hierarchical clustering

subsection, we first briefly go through these methods. Subsequently, we focus on two representative models building on these unsupervised methods, i.e., Sustain [10] and latent Dirichlet allocation [11].

2.1.1 K-Means Clustering

K-means clustering aims to directly partition the n patients into k clusters. Each patient belongs to the cluster with the nearest mean (i.e., cluster centroid) quantified by a distance metric of choice (e.g., Euclidean distance). Since searching the global minimum in clustering is computationally difficult (NP-hard), local minima are searched in the K-means algorithm via an iterative refinement approach. This usually involves two steps after giving an initial set of k centroids: (i) assignment step, assigning each data point to the cluster with the nearest centroid with the least squared Euclidean distance, and (ii) update step, recalculating means (centroids) for all data points assigned to each cluster. The two steps iteratively continue until the convergence, i.e., the assignments no longer change. More details regarding the k-means algorithm are provided in Chap. 2, Subheading 12.1. Please refer to [33–35] for representative studies using K-means for disease subtyping.

2.1.2 NMF Clustering

Nonnegative matrix factorization (NMF) is a method that implicitly performs clustering by taking advantage that complex patterns can be construed as a sum of simple parts. In essence, the input data X , is factorized into two nonnegative matrices $C \in \mathbb{R}^{p \times k}$ and $L \in \mathbb{R}^{k \times m}$, for which we refer to the component matrix and loading coefficient matrix, respectively. This method has been widely used as an effective dimensionality reduction technique in signal processing and image analysis [36]. By its nature, the L matrix can be directly used for clustering purposes, which is analogous to K-means if we impose an orthogonality constraint on the L matrix. Specifically, if $L_{kj} > L_{ij}$ for all $i \neq k$, this clusters the data point x_n into the k -th cluster. The vectors of the C matrix indicate the cluster centroids. Please refer to [32] for a representative study using NMF for disease subtyping.

2.1.3 Hierarchical Clustering

Hierarchical clustering aims to build a hierarchy of clusters, including two types of approach: agglomerative and divisive [18]. In general, the merges and splits are determined greedily and presented in a dendrogram. Similarly, a measure of dissimilarity between sets of observations is required. Most commonly, this is achieved by using an appropriate metric (e.g., Euclidean distance) and a linkage criterion that specifies the dissimilarity of sets as a function of the pairwise distances of observations. Please refer to [24, 25, 30, 37, 38] for representative studies using the hierarchical clustering for disease subtyping.

2.1.4 Representative Unsupervised Clustering Methods

Sustain [10] is an unsupervised clustering method for subtype and stage inference. Specifically, Sustain formulates the model as groups of subjects with a particular biomarker progression pattern as a subtype. The biomarker evolution of each subtype is modeled as a linear z-score model, a continuous generalization of the original event-based model [39]. Each biomarker follows a piecewise linear trajectory over a common timeframe. The key advantage of this model is that it can work with purely cross-sectional data and derive an imaging signatures of subtype and stage simultaneously.

A Bayesian latent Dirichlet allocation model [11] was proposed to extract latent AD-related atrophy factors. This probabilistic approach hypothesizes that each patient expresses one or more latent factors, and each factor is associated with distinct but possibly overlapping atrophy patterns. However, due to the nature of latent Dirichlet allocation methods, the input images have to be discretized. Moreover, this method exclusively models brain atrophy while ignoring brain enlargement. For example, larger brain volumes in basal ganglia have been associated with one subtype of schizophrenia [12].

2.2 Semi-supervised Clustering

Semi-supervised clustering methods dissect the subtle heterogeneity of interest under the principle of deriving data-driven and neurobiologically plausible subtypes (Fig. 2). In essence, these methods seek the “1-to- k ” mapping between the reference CN

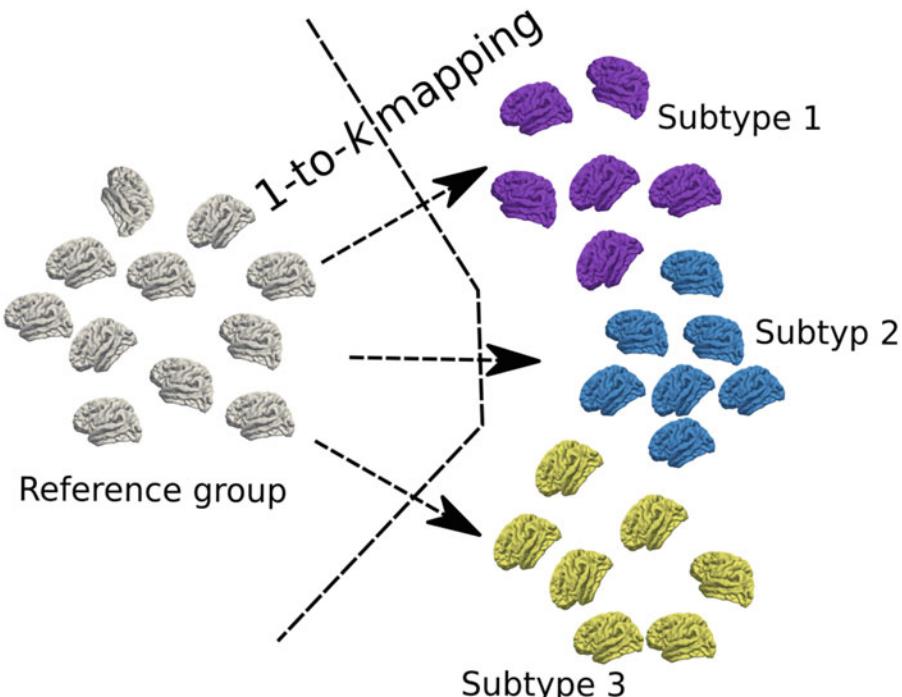


Fig. 2 Schematic diagram of semi-supervised clustering methods. Figure is adapted from [14]

group and the PT group, thereby teasing out clusters that are likely driven by distinct pathological trajectories, instead of by global similarity/dissimilarity in data, which is the core momentum of conventional unsupervised clustering methods.

In the following subsections, we briefly discuss four semi-supervised clustering methods. These methods employ different techniques to seek this “1-to- k ” mapping. In particular, CHIMERA [16] and Smile-GAN [9] utilize generative models to achieve this mapping, while HYDRA [6] and MAGIC [8] are built on top of discriminative models.

Box 1: Representative Semi-supervised Clustering Methods

The central principle of semi-supervised clustering methods is to seek the “1-to- k ” mapping from the reference domain to the patient domain.

- CHIMERA: a generative approach that leverages the coherent point drift algorithm and maps the data distribution of the CN group to the PT group, thereby enabling to subtype by the distinct k regularized transformations.
- Smile-GAN: a generative approach based on GANs to learn multiple distinct mappings by generating PT from CN. Simultaneously, a clustering model is trained interactively with mapping functions to assign PT into the corresponding subtype memberships.
- HYDRA: a discriminative approach which leverages multiple linear support vector machines to construct a polytope that clusters the patients depending on the patterns of differences between the CN group and the PT group.
- MAGIC: a generalization of HYDRA that aims to dissect disease heterogeneity at multiple imaging scales for a scale-consistent solution.

2.2.1 CHIMERA

CHIMERA employs a generative probabilistic approach, considers all samples as points in the imaging space, and infers the clusters from the transformations between the CN and PT distributions. It hypothesizes that the PT distribution can be generated from the CN distribution under k sets of transformations, each reflecting a distinct disease process.

Mathematically, the transformation \mathbf{T} is a convex combination of the k linear transformations that map a CN subject in the reference space to the target space: $\mathbf{x}_i^r \in \mathbb{R}^q \rightarrow \mathbf{x}_i^t = \mathbf{T}(\mathbf{x}_i) = \sum_{j=1}^k \xi_j \mathbf{T}_j(\mathbf{x}_i)$, where ξ_j is the probability that a PT belongs to the j -th subtype. Ideally, if the disease subtypes were distinct, ξ_j should take value 1 for the transformation

corresponding to this specific disease subtype and value 0 otherwise. At its core, the coherent point drift algorithm [40], a generative probabilistic approach, is used to estimate the transformation T . Specifically, the CN sample point is mapped to the PT domain and regarded as a centroid of a spherical Gaussian cluster, whereas the patient points are treated as independent and identically distributed data generated by a Gaussian mixture model (GMM) with equal weights for each cluster. The goal is to maximize the data likelihood during the distribution matching while also taking into account covariate confounds (e.g., age and gender). The expectation-maximization approach is adopted to optimize the resulting energy objective. Clustering inference is straightforward after the optimized transformation T_j is achieved, i.e., a patient can be assigned the subtype membership corresponding to the largest likelihood.

2.2.2 Smile-GAN

Smile-GAN is a novel generative deep learning approach based on generative adversarial networks (GAN). The reader may refer to Chap. 5 for generic information about GANs. Smile-GAN aims to learn a mapping function, f , from joint CN domain \mathcal{X} and subtype domain \mathcal{Z} to the PT domain \mathcal{Y} , by transforming CN data \mathbf{x} to different synthesized PT data $\mathbf{y}' = f(\mathbf{x}, \mathbf{z})$ that are indistinguishable from real PT data, \mathbf{y} , by the discriminator, D . Mapping function, f , is regularized for inverse consistencies, with a clustering function, $\mathcal{J} : \mathcal{Y} \rightarrow \mathcal{Z}$, trained interactively to reconstruct \mathbf{z} from synthesized PT data \mathbf{y}' . The clustering function, \mathcal{J} , can also be directly used to cluster both training and unseen test data after the training process.

More specifically, three different data distributions are denoted as $\mathbf{x} \sim p_{\text{CN}}$ (for controls), $\mathbf{y} \sim p_{\text{PT}}$ (for patients), and $\mathbf{z} \sim p_{\text{Sub}}$ (for a subtype), respectively, where $\mathbf{z} \sim p_{\text{Sub}}$ is sampled from a discrete uniform distribution and encoded as a one-hot vector with dimension K (number of clusters). Mapping function, $f : \mathcal{X} * \mathcal{Z} \rightarrow \mathcal{Y}$, and clustering function, $\mathcal{J} : \mathcal{Y} \rightarrow \mathcal{Z}$, are learned through the following training procedure (l_c denotes the cross-entropy loss):

$$f, \mathcal{J} = \arg \min_{f, \mathcal{J}} \max_D L_{\text{GAN}}(D, f) + \mu L_{\text{change}}(f) + \lambda L_{\text{cluster}}(f, \mathcal{J}) \quad (1)$$

where

$$\begin{aligned} L_{\text{GAN}}(D, f) &= \mathbb{E}_{\mathbf{y} \sim p_{\text{PT}}} [\log(D(\mathbf{y}))] \\ &\quad + \mathbb{E}_{\mathbf{z} \sim p_{\text{Sub}}, \mathbf{x} \sim p_{\text{CN}}} [1 - \log(D(f(\mathbf{x}, \mathbf{z})))] \end{aligned} \quad (2)$$

$$L_{\text{change}}(f) = \mathbb{E}_{\mathbf{x} \sim p_{\text{CN}}, \mathbf{z} \sim p_{\text{Sub}}} [|f(\mathbf{x}, \mathbf{z}) - \mathbf{x}|_1] \quad (3)$$

$$L_{\text{cluster}}(f, \mathcal{J}) = \mathbb{E}_{\mathbf{x} \sim p_{\text{CN}}, \mathbf{z} \sim p_{\text{Sub}}} [l_c(\mathbf{z}, \mathcal{J}(f(\mathbf{x}, \mathbf{z})))] \quad (4)$$

The objective consists of adversarial loss L_{GAN} , regularization terms L_{change} and L_{cluster} . Adversarial loss forces the synthesized PT data to follow similar distributions as real PT data. The discriminator D , trying to identify synthesized PT data from real PT data, attempts to maximize the loss, while the mapping f attempts to minimize against it. Both regularization terms serve to constrain the function class where the mapping function f is sampled from so that it is truly meaningful while matching the distributions. Minimization of L_{change} encourages sparsity of regions captured by f , with the assumption that only some regions are changed by disease effect. Optimizing L_{cluster} ensures that the input sub variable \mathbf{z} can be reconstructed from synthesized PT data \mathbf{y}' , so that the mutual information between \mathbf{z} and \mathbf{y}' are maximized, and distinct imaging patterns are synthesized when \mathbf{z} takes different values. Further regularization is also imposed by forcing mapping function f and clustering function g to be Lipschitz continuous. More importantly, thanks to the inverse consistencies led by L_{cluster} , function g can directly output cluster probabilities and cluster labels when given unseen test PT data.

2.2.3 HYDRA

In contrast to the generative approaches used in CHIMERA and Smile-GAN, HYDRA leverages a widely used discriminative method, i.e., support vector machines (SVM), to seek this “1-to- k ” mapping. The novelty is that HYDRA extends multiple linear SVMs to the nonlinear case in a piecewise fashion, thereby simultaneously serving for classification and clustering. Specifically, it constructs a convex polytope by combining the hyperplane from k linear SVMs, separating the CN group and the k subpopulation of the PT group. Intuitively, each face of the convex polytope can be regarded to encode each subtype, capturing a distinct disease effect.

The convex polytope is estimated by sequentially solving each linear SVM as a subproblem under the principle of the sample weighted SVM [41]. The optimization stops when the sample weights become stable, i.e., the polytope is stably established. The objective of maximizing the polytope’s margin can be summarized as

$$\begin{aligned} \min_{\{\mathbf{w}_j, b_j\}_{(j=1)}^k} & \sum_{(j=1)}^k \frac{\|\mathbf{w}_j\|_2^2}{2} + \mu \sum_{i|y_i=+1} \frac{1}{k} \max\{0, 1 - \mathbf{w}_j^T \mathbf{X}_i^T - b_j\} \\ & + \mu \sum_{i|y_i=-1} s_{i,j} \max\{0, 1 + \mathbf{w}_j^T \mathbf{X}_i^T + b_j\} \end{aligned} \quad (5)$$

where \mathbf{w}_j and b_j are the weight and bias for each hyperplane, respectively. μ is a penalty parameter on the training error, and S is the subtype membership matrix of dimension $m*k$ deciding

whether a patient sample i belongs to subtype j . The cluster membership is inferred as follows:

$$S_{i,j} = \begin{cases} 1, & j = \arg \max_j (\mathbf{w}_j^T \mathbf{X}^T + b_j) \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

2.2.4 MAGIC

MAGIC was proposed to overcome one of the main limitations that HYDRA faced. That is, a single-scale set of features (e.g., atlas-based regions of interest) may not be sufficient to derive subtle differences, compared to global demographics, disease heterogeneity, since ample evidence has shown that the brain is fundamentally composed of multi-scale structural or functional entities. To this objective, MAGIC extracts multi-scale features in a coarse-to-fine granular fashion via stochastic orthogonal projective nonnegative matrix factorization (opNMF) [42], a very effective unbiased, data-driven method for extracting biologically interpretable and reproducible feature representations. Together with these multi-scale features, HYDRA is embedded into a double-cyclic optimization procedure to yield robust and scale-consistent cluster solutions.

MAGIC encapsulates the two previous proposed methods (i.e., opNMF and HYDRA) and optimizes the clustering objective for each single-scale feature as a sub-optimization problem. To fuse the multi-scale clustering information and enforce the clusters to be scale-consistent, it adopts a double-cyclic procedure that transfers and fine-tunes the clustering polytope. Firstly, (i) inner cyclic procedure: let us remind that HYDRA decides the clusters based on the subtype membership matrix (S). MAGIC first initializes the S matrix with a specific single-scale feature set, i.e., L_i , and then the S matrix is transferred to the next set of feature set L_{i+1} until the predefined stopping criterion is achieved (i.e., the clustering solution across scales is stable). Secondly, (ii) outer cyclic procedure: the inner cyclic procedure was repeated by initializing with each single-scale feature set. Finally, to determine the final subtype assignment, we perform a consensus clustering by computing a co-occurrence matrix based on all the clustering results and then perform spectral clustering [43].

3 Application to Brain Disorders

Brain disorders and diseases affect the human brain across a wide age range. Neurodevelopmental disorders, such as autism spectrum disorders (ASD), are usually present from early childhood and affect daily functioning [44]. Psychotic disorders, such as schizophrenia, involve psychosis that is typically diagnosed for the first time in late adolescence or early adulthood [45]. Dementia and mild cognitive impairment (MCI) prevail both in late mid-life for

early-onset AD (usually 30–60 years of age) and most frequently in late-life for late-onset AD (usually over 65 years of age) [46]. Brain cancers in children and adults are heterogeneous and encompass over 100 different histological types of tumors, based on cells of origin and other histopathological features, and have substantial morbidity and mortality [47]. Ample clinical evidence encourages the stratification of the patients in these brain disorders and cancers, potentially paving the road toward individualized precision medicine.

This section collectively overviews previous work aiming to unravel imaging-derived heterogeneity in ASD, psychosis, major depressive disorders (MDD), MCI and AD, and brain cancer.

3.1 Autism Spectrum Disorder

ASD encompasses a broad spectrum of social deficits and atypical behaviors [48]. Heterogeneity of its clinical presentation has sparked massive research efforts to find subtypes to better delineate its diagnosis [49, 50]. Recent initiatives to aggregate neuroimaging data of ASD, such as the ABIDE [51] and the EU-AIMS [52], also have motivated large-scale subtyping projects using imaging signatures [53].

Different clustering methods have been applied to reveal structural brain-based subtypes, but primarily traditional techniques such as the K-means [54] or hierarchical clustering [37]. Besides structural MRI, functional MRI [55] and EEG [56] have also been popular modalities. For reasons discussed earlier, normative clustering and dimensional analyses are better suited to parse a patient population that is highly heterogeneous [57]. However, efforts in this avenue have been primitive, with only a few recent publications using cortical thickness [58]. Taken together, although more validation and replication efforts are necessary to define any reliable neuroanatomical subtypes of ASD, some convergence in findings has been noted [53]. First, most sets of ASD neuroimaging subtypes indicate a combination of both increases and decreases in imaging features compared to the CN group, instead of pointing in a uniform direction. Second, most subtypes are characterized by spatially distributed imaging patterns instead of isolated or focal patterns. Both findings emphasize the significant heterogeneity in ASD brains and the need for better stratification.

The search for subtypes in the ASD population has unique challenges. First, the early onset of ASD implies that it is heavily influenced by neurodevelopmental processes. Depending on the selected age range, the results may significantly differ. Second, ASD is more prevalent in males, with three to four male cases for one female case [59], which adds a layer of potential bias. Third, individuals with ASD often suffer psychiatric comorbidities, such as ADHD, anxiety disorders, and obsessive-compulsive disorder, among many others [60], which, if not screened carefully, can dilute or alter the true signal.

3.2 *Psychosis*

Psychosis is a medical syndrome characterized by unusual beliefs called delusions and sometimes hallucinations of visions, sounds, smells, or body sensations that are not present in reality. Symptoms, functioning, and outcomes are highly heterogeneous across individuals, leading to long-standing hypotheses of underlying brain subgroups. However, objective brain biomarkers have largely not been discovered for any psychosis diagnosis, stage, or clinically defined subgroup [61, 62]. Neuroimaging studies are also affected by brain heterogeneity [63, 64]. Recent research has thus focused on finding structural brain subtypes using unbiased statistical techniques [12, 13, 65].

Psychosis studies have mainly focused on determining subtypes by clustering brain structural data within the chronic schizophrenia population that has had the illness for years, with results demonstrating two [12, 13], three [26], and six [31] subgroups. Various clustering techniques have been used to achieve these outcomes, including conventional approaches, such as k-means, in addition to more advanced machine learning methods, such as semi-supervised learning. A limitation of the work so far has been the lack of internal or external validation. Still, in studies with robust internal validation methods using metrics that choose the optimal cluster number based on the stability of the solution (e.g., consensus clustering), subtypes cluster along the lines of the severity of brain differences.

In a recent study, with the largest sample to date ($n=671$), clustered individuals with chronic schizophrenia using HYDRA and multiple internal validation procedures were applied (i.e., cross-validation resampling, split-half reproducibility, and leave-site-out validation) [12]. A two-subtype solution was found, with one subtype demonstrating widespread reductions and the other showing the localized larger volume of the striatum that was not associated with antipsychotic use. Interestingly, there were limited associations with current psychosis symptoms in this work, but indications of associations with education and illness duration in specific subtypes.

Functional imaging has also been used to define psychosis subgroups using functional connectivity at rest [66] and effective connectivity during task performance [67]. The research commonly has relatively low sample sizes with little internal or external validation. Still, of these works, preliminary results demonstrate that clusters can follow diagnostic divisions between individuals with psychosis [67] and that specific networks (e.g., frontoparietal network) are associated with specific psychotic symptoms [67] [66]. A recent advanced deep learning approach has also revealed clinical separations along the lines of symptom severity [68]. Taken together with brain structural results, it is possible that functional imaging maps onto symptom states rather than underlying illness traits that are captured by structural imaging. Further internal and external validation work is required to investigate this hypothesis by

characterizing, comparing, and ultimately combining clustering solutions. A critical future direction will also be to conduct longitudinal studies that track individuals over time. Such research could lead the way toward clinical translation.

3.3 Major Depressive Disorder

MDD is a common, severe, and recurrent disorder, with over 300 million people affected worldwide, and is characterized by low mood, apathy, and social withdrawal, with symptoms spanning multiple domains [69]. Its vast heterogeneity is exemplified by the fact that according to DSM-5 criteria, at least 227 and up to 16,400 unique symptom presentations exist [70, 71]. The potential causes for this heterogeneity vary from divergent clinical symptom profiles to genetic etiologies and individual differences in treatment outcomes.

Despite neurobiological findings in MDD spanning cortical thickness, gray matter volume (GMV), and fractional anisotropy (FA) measures, objective brain biomarkers that can be used to diagnose and predict disease course and outcome remain elusive [71–73]. Recently, there have been efforts to identify neurobiologically based subtypes of depression using a bottom-up approach, mainly using data from resting-state fMRI [71]. Several studies [33–35] employed k-means clustering and group iterative multiple model estimation, respectively, to identify two functional connectivity subtypes, while Tokuda et al. [74] and Drysdale et al. [75] identified three and four subtypes, respectively, using nonparametric Bayesian mixture models and hierarchical clustering. These subtypes are characterized by reduced connectivity in different networks, including the default mode network (DMN), ventral attention network, and frontostriatal and limbic dysfunction. Regarding structural neuroimaging, one study has used k-means clustering on fractional anisotropy (FA) data to identify two depression subtypes. The first subtype was characterized by decreased FA in the right temporal lobe and the right middle frontal areas and was associated with an older age at onset. In contrast, the second subtype was characterized by increased FA in the left occipital lobe and was associated with a younger age at onset [76].

Current research in the identification of brain subtypes in MDD has produced results that are promising but confounded by methodological and design limitations. While some studies have shown clinical promises such as predicting higher depressive symptomatology and lower sustenance of positive mood [34, 35], depression duration [33], and TMS therapy response [75], they are confounded by limitations such as relatively small sample sizes; nuisance variances such as age, gender, and common ancestry; lack of external validation; and lack of statistical significance testing of identified clusters. Furthermore, there has been a lack of ambition in the use of novel clustering techniques. Clustering based on

structural neuroimaging is limited compared to other disease entities and is an avenue that future research should consider. Future studies should also aim to perform longitudinal clustering to elucidate the stability of identified brain subtypes over time and examine their utility in predicting disease outcomes.

3.4 MCI and AD

AD, along with its prodromal stage presenting MCI, is the most common neurodegenerative disease, affecting millions across the globe. Although a plethora of imaging studies have derived AD-related imaging signatures, most studies ignored the heterogeneity in AD. Recently, there has been a developing body of effort to derive imaging signatures of AD that are heterogeneity-aware (i.e., subtypes) [7–11].

Most previous studies leveraged unsupervised clustering methods such as Sustain [10], NMF [32], latent Dirichlet allocation [11], and hierarchical clustering [24, 25, 30, 38]. Other papers [6, 9, 20, 77, 78] utilized semi-supervised clustering methods. Due to the variabilities of the choice of databases and methodologies and the lack of ground truth in the context of clustering, the reported number of clusters and the subtypes' neuroanatomical patterns differ and cannot be directly compared. The targeted heterogeneous population of study also varies across papers. For instance, [6] focused on dissecting the neuroanatomical heterogeneity for AD patients, while [77] included AD plus MCI and [20] studied MCI only. However, some common subtypes were found in different studies. First, a subtype showing a typical diffuse atrophy pattern over the entire brain was witnessed in several studies [6, 8–10, 22, 27, 29, 30, 32, 38, 77]. Another subtype demonstrating nearly normal brain anatomy was robustly identified [8, 9, 16, 20, 22, 24, 25, 29, 30]. Moreover, studies [8, 9, 29, 30, 77] also reported one subtype showing atypical AD patterns (i.e., hippocampus or medial temporal lobe spared).

Though these methods enabled a better understanding of heterogeneity in AD, there are still limitations and challenges. First, due to demographic variations and the existence of comorbidities, it is not guaranteed that models cluster the data based on variations of the pathology of interest. Semi-supervised methods might tackle this problem to some extent, but more careful sample selection and further study with longitudinal data may ensure disease specificity. Second, spatial differences and temporal changes may simultaneously contribute to subtypes derived through clustering methods. Third, subtypes captured from neuroimaging data alone bring limited insight into disease treatments, thereby a joint study of neuroimaging and genetic heterogeneity may provide greater clinical value [14, 79].

3.5 Brain Cancer

Brain tumors, such as glioblastoma (GBM), exhibit extensive inter- and intra-tumor heterogeneity, diffuse infiltration, and invasiveness of various immune and stromal cell populations, which pose diagnostic and prognostic challenges, and render the standard therapies futile [80]. Deciphering the underlying heterogeneity of brain tumors, which arises from genomic instability of these tumors, plays a key role in understanding and predicting the course of tumor progression and its response to the standard therapies, thereby designing effective therapies targeted at aberrant genetic alterations [81, 82]. Medical imaging noninvasively portrays the phenotypic differences of brain tumors and their microenvironment caused by molecular activities of tumors on a macroscopic scale [83, 84]. It has the potential to provide readily accessible and surrogate biomarkers of particular genomic alterations, predict response to therapy, avoid risks of tumor biopsy or inaccurate diagnosis due to sampling errors, and ultimately develop personalized therapies to improve patient outcomes. An imaging subtype of brain tumors may provide a wealth of information about the tumor, including distinct molecular pathways [85, 86].

Recent studies on radiomic analysis of multiparametric MRI (mpMRI) scans provide evidence of distinct phenotypic presentation of brain tumors associated with specific molecular characteristics. These studies propose that quantification of tumor morphology, texture, regional microvasculature, cellular density, or microstructural properties can map to different imaging subtypes. In particular, one study [87] discovered three distinct clusters of GBM subtypes through unsupervised clustering of these features, with significant differences in survival probabilities and associations with specific molecular signaling pathways. These imaging subtypes, namely solid, irregular, and rim-enhancing, were significantly linked to different clinical outcomes and molecular characteristics, including isocitrate dehydrogenase-1, O6-methylguanine-DNA methyltransferase, epidermal growth factor receptor variant III, and transcriptomic molecular subtype composition.

These studies have offered new insights into the characterization of tumor heterogeneity on both microscopic, i.e., histology and molecular, and macroscopic, i.e., imaging levels, consequently providing a more comprehensive understanding of the tumor aggressiveness and patient prognosis, and ultimately, the development of personalized treatments.

4 Conclusion

Taken together, these novel clustering algorithms tailored for high-resolution yet highly variable neuroimaging datasets have demonstrated a broad utility in disease subtyping across many neurological and psychiatric conditions. Simultaneously, cautions need to be

taken in order not to overclaim the biological importance of subtypes, since all clustering methods find patterns in data, even if such patterns don't have a meaningful underlying biological correlate [88]. External validations are necessary. For instance, evidence of post hoc evaluations, e.g., a difference in clinical variables or genetic architectures, can support the biological relevance of identified neuroimaging-based subtypes [14]. Moreover, good practices such as split-sample analysis, permutation tests [12], and comparison to the guideline of semi-simulated experiments [8] discern the robustness of the subtypes. As dataset sizes and imaging resolution improve over time, unique computational challenges are expected to appear, along with unique opportunities to further refine our methodologies to decipher the diversity of brain diseases.

Acknowledgements

This work was supported, in part, by NIH grants R01NS042645, U01AG068057, R01MH112070, and RF1AG054409.

References

- Murray ME, Graff-Radford NR, Ross OA, Petersen RC, Duara R, Dickson DW (2011) Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol* 10(9):785–796. [https://doi.org/10.1016/S1474-4422\(11\)70156-9](https://doi.org/10.1016/S1474-4422(11)70156-9)
- Noh Y, Jeon S, Lee JM, Seo SW, Kim GH, Cho H, Ye BS, Yoon CW, Kim HJ, Chin J, Park KH, Heilman KM, Na DL (2014) Anatomical heterogeneity of Alzheimer disease: based on cortical thickness on MRIs. *Neurology* 83(21):1936–1944. <https://doi.org/10.1212/WNL.0000000000001003>
- Whitwell JL, Petersen RC, Negash S, Weigand SD, Kantarci K, Iynik RJ, Knopman DS, Boeve BF, Smith GE, Jack CR (2007) Patterns of atrophy differ among specific subtypes of mild cognitive impairment. *Arch Neurol* 64(8):1130–1138. <https://doi.org/10.1001/archneur.64.8.1130>
- Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, Bartsch AJ, Jbabdi S, Sotiroopoulos SN, Andersson JLR, Griffanti L, Douaud G, Okell TW, Weale P, Dragou I, Garratt S, Hudson S, Collins R, Jenkinson M, Matthews PM, Smith SM (2016) Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* 19(11):1523–1536. <https://doi.org/10.1038/nn.4393>
- Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, Jack CR, Jagust WJ, Shaw LM, Toga AW, Trojanowski JQ, Weiner MW (2010) Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology* 74(3):201–209. <https://doi.org/10.1212/WNL.0b013e3181cb3e25>
- Varol E, Sotiras A, Davatzikos C (2017) HYDRA: revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. *NeuroImage* 145:346–364. <https://doi.org/10.1016/j.neuroimage.2016.02.041>
- Vogel JW, Young AL, Oxtoby NP, Smith R, Ossenkoppele R, Strandberg OT, La Joie R, Aksman LM, Grothe MJ, Iturria-Medina Y, Pontecorvo MJ, Devous MD, Rabinovici GD, Alexander DC, Lyoo CH, Evans AC, Hansson O (2021) Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med* 27(5):871–881. <https://doi.org/10.1038/s41591-021-01309-6>
- Wen J, Varol E, Sotiras A, Yang Z, Chand GB, Erus G, Shou H, Abdulkadir A, Hwang G, Dwyer DB, Pigoni A, Dazzan P, Kahn RS, Schnack HG, Zanetti MV, Meisenzahl E, Busatto GF, Crespo-Facorro B, Rafael RG, Pantelis C, Wood SJ, Zhuo C, Shinohara RT, Fan Y, Gur RC, Gur RE, Satterthwaite TD, Koutsouleris N, Wolf DH, Davatzikos C, Alzheimer's disease neuroimaging initiative

- (2021) Multi-scale semi-supervised clustering of brain images: deriving disease subtypes. *Med Image Anal* 75:102304. <https://doi.org/10.1016/j.media.2021.102304>
9. Yang Z, Nasrallah IM, Shou H, Wen J, Doshi J, Habes M, Erus G, Abdulkadir A, Resnick SM, Albert MS, Maruff P, Fripp J, Morris JC, Wolk DA, Davatzikos C, iSTAGING Consortium, Baltimore Longitudinal Study of Aging (BLSA), Alzheimer's Disease Neuroimaging Initiative (ADNI) (2021) A deep learning framework identifies dimensional representations of Alzheimer's disease from brain structure. *Nature Communications* 12(1):7065. <https://doi.org/10.1038/s41467-021-26703-z>
 10. Young A, Marinescu R, Oxtoby N, Bocchetta M, Yong K, Firth N, Cash D, Thomas D, Moore K, Cardoso MJ, Swieten J, Borroni B, Galimberti D, Masellis M, Tartaglia M, Rowe J, Graff C, Tagliavini F, Frisoni G, Alexander D (2018) Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nat Commun*. <https://doi.org/10.1038/s41467-018-05892-0>
 11. Zhang X, Mormino E, Sun N, Sperling R, Sabuncu M, Yeo BT (2016) Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. *Proc Natl Acad Sci* 113:E6535–E6544. <https://doi.org/10.1073/pnas.1611073113>
 12. Chand GB, Dwyer DB, Erus G, Sotiras A, Varol E, Srinivasan D, Doshi J, Pomponio R, Pigoni A, Dazzan P, Kahn RS, Schnack HG, Zanetti MV, Meisenzahl E, Busatto GF, Crespo-Facorro B, Pantelis C, Wood SJ, Zhuo C, Shinohara RT, Shou H, Fan Y, Gur RC, Gur RE, Satterthwaite TD, Koutsouleris N, Wolf DH, Davatzikos C (2020) Two distinct neuroanatomical subtypes of schizophrenia revealed using machine learning. *Brain* 143(3):1027–1038. <https://doi.org/10.1093/brain/awaa025>
 13. Dwyer DB, Cabral C, Kambeitz-Ilankovic L, Sanfelici R, Kambeitz J, Calhoun V, Falkai P, Pantelis C, Meisenzahl E, Koutsouleris N (2018) Brain subtyping enhances the neuroanatomical discrimination of schizophrenia. *Schizophr Bull* 44(5):1060–1069. <https://doi.org/10.1093/schbul/sby008>
 14. Wen J, Fu CHY, Tosun D, Veturi Y, Yang Z, Abdulkadir A, Mamourian E, Srinivasan D, Ioanna S, Ashish S, Bao J, Erus G, Shou H, Habes M, Doshi J, Varol E, Mackin SR, Sotiras A, Fan Y, Saykin AJ, Sheline YI, Shen L, Ritchie MD, Wolk DA, Albert M, Resnick SM, Davatzikos C (2022) Characterizing heterogeneity in neuroimaging, cognition, clinical symptomatology, and genetics among patients with late-life depression. *JAMA Psychiatry*. <https://doi.org/10.1001/jamapsychiatry.2022.0020>
 15. Yang Z DC Wen P (2022) Surreal-gan:semi-supervised representation learning via GAN for uncovering heterogeneous disease-related imaging patterns. *ICLR* 2022. <https://openreview.net/forum?id=nf3A0WZsXS5>
 16. Dong A, Honnorat N, Gaonkar B, Davatzikos C (2016) CHIMERA: clustering of heterogeneous disease effects via distribution matching of imaging patterns. *IEEE Trans Med Imaging* 35(2):612–621. <https://doi.org/10.1109/TMI.2015.2487423>
 17. Hamerly G, Elkan C (2004) Learning the k in k -means. In: Thrun S, Saul LK, Schölkopf B (eds) *Advances in neural information processing systems*, vol 16. MIT Press, Cambridge, MA, pp 281–288. <http://papers.nips.cc/paper/2526-learning-the-k-in-k-means.pdf>
 18. Day WHE, Edelsbrunner H (1984) Efficient algorithms for agglomerative hierarchical clustering methods. *J Classif* 1(1):7–24. <https://doi.org/10.1007/BF01890115>
 19. Davatzikos C (2019) Machine learning in neuroimaging: progress and challenges. *NeuroImage* 197:652–656. <https://doi.org/10.1016/j.neuroimage.2018.10.003>
 20. Ezzati A, Zammit AR, Habeck C, Hall CB, Lipton RB, Alzheimer's Disease Neuroimaging Initiative (2020) Detecting biological heterogeneity patterns in ADNI amnestic mild cognitive impairment based on volumetric MRI. *Brain Imaging Behav* 14(5):1792–1804. <https://doi.org/10.1007/s11682-019-00115-6>
 21. Honnorat N, Dong A, Meisenzahl-Lechner E, Koutsouleris N, Davatzikos C (2019) Neuroanatomical heterogeneity of schizophrenia revealed by semi-supervised machine learning methods. *Schizophr Res* 214:43–50. <https://doi.org/10.1016/j.schres.2017.12.008>
 22. Jung NY, Seo SW, Yoo H, Yang JJ, Park S, Kim YJ, Lee J, Lee JS, Jang YK, Lee JM, Kim ST, Kim S, Kim EJ, Na DL, Kim HJ (2016) Classifying anatomical subtypes of subjective memory impairment. *Neurobiol Aging* 48:53–60. <https://doi.org/10.1016/j.neurobiolaging.2016.08.010>
 23. Lubeiro A, Rueda C, Hernández JA, Sanz J, Sarramea F, Molina V (2016) Identification of two clusters within schizophrenia with different structural, functional and clinical characteristics. *Prog Neuro-Psychopharmacol Biol Psychiatry* 64:79–86. <https://doi.org/10.1016/j.pnpbp.2015.06.015>

24. Nettiksimmons J, DeCarli C, Landau S, Beckett L (2014) Biological heterogeneity in ADNI amnestic mild cognitive impairment. *Alzheimer's Dementia* 10(5):511–521.e1. <https://doi.org/10.1016/j.jalz.2013.09.003>
25. Ota K, Oishi N, Ito K, Fukuyama H (2016) Prediction of Alzheimer's disease in amnestic mild cognitive impairment subtypes: Stratification based on imaging biomarkers. *J Alzheimer's Disease*. <https://doi.org/10.3233/JAD-160145>
26. Pan Y, Pu W, Chen X, Huang X, Cai Y, Tao H, Xue Z, Mackinlay M, Limongi R, Liu Z, Palaniyappan L (2020) Morphological profiling of schizophrenia: cluster analysis of MRI-based cortical thickness data. *Schizophr Bull* 46(3): 623–632. <https://doi.org/10.1093/schbul/sbz112>
27. Park JY, Na HK, Kim S, Kim H, Seo S, Na D, Han C, Seong JK (2017) Robust identification of Alzheimer's disease subtypes based on cortical atrophy patterns. *Sci Rep* 7:43270. <https://doi.org/10.1038/srep43270>
28. Planchuelo-Gomez A, Lubeiro A, Nunez-Novo P, Gomez-Pilar J, de Luis-Garcia R, Del Valle P, Martin-Santiago O, Perez-Escudero A, Molina V (2020) Identificacion of MRI-based psychosis subtypes: replication and refinement. *Progr Neuro-Psychopharmacol Biol Psychiatry* 100:109907. <https://doi.org/10.1016/j.pnpbp.2020.109907>
29. Poulakis K, Ferreira D, Pereira JB, Smedby O, Vemuri P, Westman E (2020) Fully Bayesian longitudinal unsupervised learning for the assessment and visualization of AD heterogeneity and progression. *Aging* 12(13): 12622–12647
30. Poulakis K, Pereira JB, Mecocci P, Vellas B, Tsolaki M, Kloszewska I, Soininen H, Lovestone S, Simmons A, Wahlund LO, Westman E (2018) Heterogeneous patterns of brain atrophy in Alzheimer's disease. *Neurobiol Aging* 65:98–108. <https://doi.org/10.1016/j.neurobiolaging.2018.01.009>
31. Sugihara G, Oishi N, Son S, Kubota M, Takahashi H, Murai T (2017) Distinct Patterns of cerebral cortical thinning in schizophrenia: a neuroimaging data-driven approach. *Schizophr Bull* 43(4):900–906. <https://doi.org/10.1093/schbul/sbw176>
32. ten Kate M, Dicks E, Visser PJ, van der Flier WM, Teunissen CE, Barkhof F, Scheltens P, Tijms BM, Alzheimer's Disease Neuroimaging Initiative (2018) Atrophy subtypes in prodromal Alzheimer's disease are associated with cognitive decline. *Brain* 141(12):3443–3456. <https://doi.org/10.1093/brain/awy264>
33. Feder S, Sundermann B, Wersching H, Teuber A, Kugel H, Teismann H, Heindel W, Berger K, Pfeiderer B (2017) Sample heterogeneity in unipolar depression as assessed by functional connectivity analyses is dominated by general disease effects. *J Affective Disord* 222:79–87. <https://doi.org/10.1016/j.jad.2017.06.055>
34. Price RB, Gates K, Kraynak TE, Thase ME, Siegle GJ (2017) Data-driven subgroups in depression derived from directed functional connectivity paths at rest. *Neuropsychopharmacology* 42(13):2623–2632. <https://doi.org/10.1038/npp.2017.97>
35. Price RB, Lane S, Gates K, Kraynak TE, Horner MS, Thase ME, Siegle GJ (2017) Parsing Heterogeneity in the brain connectivity of depressed and healthy adults during positive mood. *Biolo Psychiatry* 81(4):347–357. <https://doi.org/10.1016/j.biopsych.2016.06.023>
36. Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In: *Proc. Neural Information Processing Systems*, p 7
37. Hong SJ, Valk SL, Di Martino A, Milham MP, Bernhardt BC (2018) Multidimensional neuroanatomical subtyping of autism spectrum disorder. *Cerebral Cortex* (New York, NY: 1991) 28(10):3578–3588. <https://doi.org/10.1093/cercor/bhx229>
38. Jeon S, Kang JM, Seo S, Jeong HJ, Funck T, Lee SY, Park KH, Lee YB, Yeon BK, Ido T, Okamura N, Evans AC, Na DL, Noh Y (2019) Topographical heterogeneity of Alzheimer's disease based on MR imaging, Tau PET, and amyloid PET. *Front Aging Neurosci* 11:211. <https://doi.org/10.3389/fnagi.2019.00211>
39. Young AL, Oxtoby NP, Daga P, Cash DM, Fox NC, Ourselin S, Schott JM, Alexander DC, Alzheimer's Disease Neuroimaging Initiative (2014) A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* 137(Pt 9):2564–2577. <https://doi.org/10.1093/brain/awu176>
40. Myronenko A SX (2010) Point set registration: coherent point drift. *IEEE Trans Pattern Anal Mach Intell* 32:2262–2275. <https://doi.org/10.1109/TPAMI.2010.46>
41. Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol* 2(3):1–27. <https://doi.org/10.1145/1961189.1961199>
42. Sotiras A, Resnick SM, Davatzikos C (2015) Finding imaging patterns of structural covariance via non-negative matrix factorization. *NeuroImage* 108:1–16. <https://doi.org/10.1016/j.neuroimage.2014.11.045>

43. Ng A, Jordan M, Weiss Y (2002) On spectral clustering: analysis and an algorithm. In: Dietterich T, Becker S, Ghahramani Z (eds) Advances in neural information processing systems, vol 14. MIT Press, Cambridge, MA. <https://proceedings.neurips.cc/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf>
44. Faras H, Al AN, Tidmarsh L (2010) Autism spectrum disorders. Ann Saudi Med 30(4): 295–300. <https://doi.org/10.4103/0256-4947.65261>
45. Gottesman II, Shields J, Hanson DR (1982) Schizophrenia. CUP Archive. Google-Books-ID: eoA6AAAAIAAJ
46. Mucke L (2009) Alzheimer's disease. Nature 461(7266):895–897. <https://doi.org/10.1038/461895a>
47. Ostrom QT, Adel Fahmideh M, Cote DJ, Muskens IS, Schraw JM, Scheurer ME, Bondy ML (2019) Risk factors for childhood and adult primary brain tumors. Neuro-Oncology 21(11):1357–1375. <https://doi.org/10.1093/neuonc/noz123>
48. Masi A, DeMayo MM, Glozier N, Guastella AJ (2017) An overview of autism spectrum disorder, heterogeneity and treatment options. Neurosci Bull 33(2):183–193. <https://doi.org/10.1007/s12264-017-0100-y>
49. Lord C, Jones RM (2012) Annual research review: re-thinking the classification of autism spectrum disorders. J Child Psychol Psychiatry Allied Disciplines 53(5):490–509. <https://doi.org/10.1111/j.1469-7610.2012.02547.x>
50. Wolfers T, Floris DL, Dinga R, van Rooij D, Isakoglou C, Kia SM, Zabihi M, Llera A, Chowdanayaka R, Kumar VJ, Peng H, Laidi C, Batalle D, Dimitrova R, Charman T, Loth E, Lai MC, Jones E, Baumeister S, Moessnang C, Banaschewski T, Ecker C, Dumas G, O'Muircheartaigh J, Murphy D, Buitelaar JK, Marquand AF, Beckmann CF (2019) From pattern classification to stratification: towards conceptualizing the heterogeneity of Autism Spectrum Disorder. Neurosci Biobehav Rev 104:240–254. <https://doi.org/10.1016/j.neubiorev.2019.07.010>
51. Di Martino A, O'Connor D, Chen B, Alaerts K, Anderson JS, Assaf M, Balsters JH, Baxter L, Beggiato A, Bernaerts S, Blanken LME, Bookheimer SY, Braden BB, Byrge L, Castellanos FX, Dapretto M, Delorme R, Fair DA, Fishman I, Fitzgerald J, Gallagher L, Keehn RJ, Kennedy DP, Lainhart JE, Luna B, Mostofsky SH, Müller RA, Nebel MB, Nigg JT, O'Hearn K, Solomon M, Toro R, Vaidya CJ, Wenderoth N, White T, Craddock RC, Lord C, Leventhal B, Milham MP (2017) Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Sci Data 4:170010. <https://doi.org/10.1038/sdata.2017.10>
52. Loth E, Charman T, Mason L, Tillmann J, Jones EJH, Wooldridge C, Ahmad J, Auyeung B, Brogna C, Ambrosino S, Banaschewski T, Baron-Cohen S, Baumeister S, Beckmann C, Brammer M, Brandeis D, Bölte S, Bourgeron T, Bours C, de Brujin Y, Chakrabarti B, Crawley D, Cornelissen I, Acqua FD, Dumas G, Durston S, Ecker C, Faulkner J, Frouin V, Garces P, Goyard D, Hayward H, Ham LM, Hipp J, Holt RJ, Johnson MH, Isaksson J, Kundu P, Lai MC, D'ardhuy XL, Lombardo MV, Lythgoe DJ, Mandl R, Meyer-Lindenberg A, Moessnang C, Mueller N, O'Dwyer L, Oldehinkel M, Oranje B, Pandina G, Persico AM, Ruigrok ANV, Ruggeri B, Sabet J, Sacco R, Cáceres ASJ, Simonoff E, Toro R, Tost H, Waldman J, Williams SCR, Zwiers MP, Spooren W, Murphy DGM, Buitelaar JK (2017) The EU-AIMS Longitudinal European Autism Project (LEAP): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders. Mol Autism 8:24. <https://doi.org/10.1186/s13229-017-0146-8>
53. Hong SJ, Vogelstein JT, Gozzi A, Bernhardt BC, Yeo BTT, Milham MP, Di Martino A (2020) Toward neurosubtypes in Autism. Biol Psychiatry 88(1):111–128. <https://doi.org/10.1016/j.biopsych.2020.03.022>
54. Chen H, Uddin LQ, Guo X, Wang J, Wang R, Wang X, Duan X, Chen H (2019) Parsing brain structural heterogeneity in males with autism spectrum disorder reveals distinct clinical subtypes. Hum Brain Mapp 40(2): 628–637. <https://doi.org/10.1002/hbm.24400>
55. Easson AK, Fatima Z, McIntosh AR (2019) Functional connectivity-based subtypes of individuals with and without Autism Spectrum Disorder. Netw Neurosci 3(2):344–362. https://doi.org/10.1162/netn_a_00067
56. Duffy FH, Als H (2019) Autism, spectrum or clusters? An EEG coherence study. BMC Neurol 19(1):27. <https://doi.org/10.1186/s12883-019-1254-1>
57. Tang S, Sun N, Floris DL, Zhang X, Di Martino A, Yeo BTT (2020) Reconciling dimensional and categorical models of autism heterogeneity: a brain connectomics and behavioral study. Biol Psychiatry 87(12): 1071–1082. <https://doi.org/10.1016/j.biopsych.2019.11.009>

58. Zabih M, Floris DL, Kia SM, Wolfers T, Tillmann J, Arenas AL, Moessnang C, Banaschewski T, Holt R, Baron-Cohen S, Loth E, Charman T, Bourgeron T, Murphy D, Ecker C, Buitelaar JK, Beckmann CF, Marquand A, EU-AIMS LEAP Group (2020) Fractionating autism based on neuro-anatomical normative modeling. *Transl Psychiatry* 10(1):384. <https://doi.org/10.1038/s41398-020-01057-0>
59. Loomes R, Hull L, Mandy WPL (2017) What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *J Am Acad Child Adolesc Psychiatry* 56(6):466–474. <https://doi.org/10.1016/j.jaac.2017.03.013>
60. Gillberg C, Fennell E (2014) Autism plus versus autism pure. *J Autism Dev Disord* 44(12): 3274–3276. <https://doi.org/10.1007/s10803-014-2163-1>
61. Hajima SV, Van Haren N, Cahn W, Koolschijn PCMP, Hulshoff Pol HE, Kahn RS (2013) Brain volumes in schizophrenia: a meta-analysis in over 18000 subjects. *Schizophr Bull* 39(5): 1129–1138. <https://doi.org/10.1093/schbul/sbs118>
62. Pantelis C, Velakoulis D, McGorry PD, Wood SJ, Suckling J, Phillips LJ, Yung AR, Bullmore ET, Brewer W, Soulsby B, Desmond P, McGuire PK (2003) Neuroanatomical abnormalities before and after onset of psychosis: a cross-sectional and longitudinal MRI comparison. *Lancet* (London, England) 361(9354):281–288. [https://doi.org/10.1016/S0140-6736\(03\)12323-9](https://doi.org/10.1016/S0140-6736(03)12323-9)
63. Abi-Dargham A, Horga G (2016) The search for imaging biomarkers in psychiatric disorders. *Nat Med* 22(11):1248–1255. <https://doi.org/10.1038/nm.4190>
64. Insel T, Cuthbert B (2015) Medicine. brain disorders? precisely. *Science* (New York, NY) 348:499–500. <https://doi.org/10.1126/science.aab2358>
65. Kaczkurkin AN, Moore TM, Sotiras A, Xia CH, Shinohara RT, Satterthwaite TD (2020) Approaches to defining common and dissociable neurobiological deficits associated with psychopathology in youth. *Biol Psychiatry* 88(1):51–62. <https://doi.org/10.1016/j.biopsych.2019.12.015>
66. Yang Z, Xu Y, Xu T, Hoy CW, Handwerker DA, Chen G, Northoff G, Zuo XN, Bandettini PA (2014) Brain network informed subject community detection in early-onset schizophrenia. *Sci Rep* 4:5549. <https://doi.org/10.1038/srep05549>
67. Brodersen KH, Deserno L, Schlagenhauf F, Lin Z, Penny WD, Buhmann JM, Stephan KE (2014) Dissecting psychiatric spectrum disorders by generative embedding. *NeuroImage Clin* 4:98–111. <https://doi.org/10.1016/j.nic.2013.11.002>
68. Yan W, Zhao M, Fu Z, Pearson GD, Sui J, Calhoun VD (2022) Mapping relationships among schizophrenia, bipolar and schizoaffective disorders: a deep classification and clustering framework using fMRI time series. *Schizophr Res* 245:141–150. <https://doi.org/10.1016/j.schres.2021.02.007>
69. World Health Organization (1992) The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. Tech. rep., World Health Organization, p 362. <https://apps.who.int/iris/handle/10665/37958>. ISBN: 9787117019576
70. Fried EI, Nesse RM (2015) Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. *J Affective Disord* 172:96–102. <https://doi.org/10.1016/j.jad.2014.10.010>
71. Lynch CJ, Gunning FM, Liston C (2020) Causes and consequences of diagnostic heterogeneity in depression: paths to discovering novel biological depression subtypes. *Biol Psychiatry* 88(1):83–94. <https://doi.org/10.1016/j.biopsych.2020.01.012>
72. Buch AM, Liston C (2021) Dissecting diagnostic heterogeneity in depression by integrating neuroimaging and genetics. *Neuropsychopharmacology* 46(1):156–175. <https://doi.org/10.1038/s41386-020-00789-3>
73. Rajkowska G, Miguel-Hidalgo JJ, Wei J, Dilley G, Pittman SD, Meltzer HY, Overholser JC, Roth BL, Stockmeier CA (1999) Morphometric evidence for neuronal and glial prefrontal cell pathology in major depression. *Biol Psychiatry* 45(9):1085–1098. [https://doi.org/10.1016/s0006-3223\(99\)00041-4](https://doi.org/10.1016/s0006-3223(99)00041-4)
74. Tokuda T, Yoshimoto J, Shimizu Y, Okada G, Takamura M, Okamoto Y, Yamawaki S, Doya K (2018) Identification of depression subtypes and relevant brain regions using a data-driven approach. *Sci Rep* 8:14082. <https://doi.org/10.1038/s41598-018-32521-z>
75. Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, Fethko RN, Zebley B, Oathes DJ, Etkin A, Schatzberg AF, Sudheimer K, Keller J, Mayberg HS, Gunning FM, Alexopoulos GS, Fox MD, Pascual-Leone A, Voss HU, Casey BJ, Dubin MJ, Liston C (2017) Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med* 23(1):28–38. <https://doi.org/10.1038/nm.4246>

76. Cheng Y, Xu J, Yu H, Nie B, Li N, Luo C, Li H, Liu F, Bai Y, Shan B, Xu L, Xu X (2014) Delineation of early and later adult onset depression by diffusion tensor imaging. PLoS ONE 9(11). <https://doi.org/10.1371/journal.pone.0112307>
77. Dong A, Toledo JB, Honnorat N, Doshi J, Varol E, Sotiras A, Wolk D, Trojanowski JQ, Davatzikos C, for the Alzheimer's Disease Neuroimaging Initiative (2017) Heterogeneity of neuroanatomical patterns in prodromal Alzheimer's disease: links to cognition, progression and biomarkers. Brain 140(3):735–747. <https://doi.org/10.1093/brain/aww319>
78. Filipovich R, Resnick SM, Davatzikos C (2012) JointMMCC: joint maximum-margin classification and clustering of imaging data. IEEE Trans Med Imaging 31(5):1124–1140. <https://doi.org/10.1109/TMI.2012.2186977>
79. Chand GB, Singhal P, Dwyer DB, Wen J, Erus G, Doshi J, Srinivasan D, Mamourian E, Varol E, Sotiras A, Hwang G (2022) Two schizophrenia imaging signatures and their associations with cognition, psychopathology, and genetics in the general population. <https://doi.org/10.1101/2022.01.07.22268854>
80. Andrea S, Inmaculada S, Sara P, Anestis T, Peter C, John M, Christina C, Colin W, Simon T (2013) Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proc Natl Acad Sci USA 110: 4009–4014. <https://doi.org/10.1073/pnas.1219747110>
81. Akhavan D, Alizadeh D, Wang D, Weist MR, Shepphard JK, Brown CE (2019) CAR T cells for brain tumors: lessons learned and road ahead. Immunol Rev 290(1):60–84. <https://doi.org/10.1111/imr.12773>
82. Qazi MA, Vora P, Venugopal C, Sidhu SS, Moffat J, Swanton C, Singh SK (2017) Intratumoral heterogeneity: pathways to treatment resistance and relapse in human glioblastoma. Ann Oncol 28(7):1448–1456. <https://doi.org/10.1093/annonc/mdx169>
83. Davatzikos C, Sotiras A, Fan Y, Habes M, Erus G, Rathore S, Bakas S, Chitalia R, Gastounioti A, Kontos D (2019) Precision diagnostics based on machine learning-derived imaging signatures. Magn Reson Imaging 64: 49–61. <https://doi.org/10.1016/j.mri.2019.04.012>
84. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: images are more than pictures, they are data. Radiology 278(2):563–577. <https://doi.org/10.1148/radiol.2015151169>
85. Fathi Kazerooni A, Bakas S, Saligheh Rad H, Davatzikos C (2020) Imaging signatures of glioblastoma molecular characteristics: a radiogenomics review. J Magn Reson Imaging 52(1):54–69. <https://doi.org/10.1002/jmri.26907>
86. Gevaert O, Mitchell LA, Achrol AS, Xu J, Echegaray S, Steinberg GK, Cheshier SH, Napel S, Zaharchuk G, Plevritis SK (2014) Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. Radiology 273(1):168–174. <https://doi.org/10.1148/radiol.14131731>
87. Rathore S, Akbari H, Rozycski M, Abdullah KG, Nasrallah MP, Binder ZA, Davuluri RV, Lustig RA, Dahmane N, Bilello M, O'Rourke DM, Davatzikos C (2018) Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. Sci Rep 8(1):5087. <https://doi.org/10.1038/s41598-018-22739-2>
88. Altman N, Krzywinski M (2017) Clustering. Nat Methods 14(6):545–546. <https://doi.org/10.1038/nmeth.4299>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made. The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 17

Data-Driven Disease Progression Modeling

Neil P. Oxtoby

Abstract

Intense debate in the neurology community before 2010 culminated in hypothetical models of Alzheimer's disease progression: a pathophysiological cascade of biomarkers, each dynamic for only a segment of the full disease timeline. Inspired by this, data-driven disease progression modeling emerged from the computer science community with the aim to reconstruct neurodegenerative disease timelines using data from large cohorts of patients, healthy controls, and prodromal/at-risk individuals. This chapter describes selected highlights from the field, with a focus on utility for understanding and forecasting of disease progression.

Key words Disease progression, Disease understanding, Forecasting, Cross-sectional data, Longitudinal data, Disease timelines, Disease trajectories, Subtyping, Biomarkers

1 Introduction

Chronic progressive diseases are a major drain on social and economic resources. Many of these diseases have no treatments and no cure. In particular, age-related chronic diseases such as neurodegenerative diseases of the brain are a global healthcare pandemic-in-waiting as most of the world's population is living ever longer. A key example is Alzheimer's disease—the leading cause of dementia—but there are numerous other conditions that cause abnormal deterioration of brain tissue, leading to loss of cognitive performance, bodily function, independence, and ultimately death. Despite the increasing socioeconomic burden, neurodegenerative disease research has made impressive progress in the past decade, driven largely by the availability of large observational datasets and the computational analyses this enables.

Understanding neurodegenerative diseases is vital if they are to be managed, or even cured, but our understanding remains poor despite impressive progress in recent years. This poor understanding can be attributed to the many challenges of neurodegenerative diseases: no well-defined time axis due in part to heterogeneity in onset/speed/presentation, and censoring/attrition especially in

later stages as patients deteriorate. These challenges, coupled with intense debate in the neurology community (hypothetical models [1, 2]) and increasing availability of data, piqued the interest of computational researchers aiming to provide quantitative answers to the mysteries of neurodegenerative diseases. This has ranged from vanilla off-the-shelf machine learning approaches through to more holistic statistical modeling approaches, the most advanced of which is data-driven disease progression modeling (D^3PM).

D^3PMs are defined by two key features: (1) they simultaneously reconstruct the disease timeline and estimate the quantitative disease signature/trajecory along this timeline; and (2) they are directly informed by observed data. D^3PMs strike a balance between pure unsupervised learning, which requires truly big data, and traditional longitudinal modeling, which relies on a well-defined temporal axis—neither of which are available in neurodegenerative diseases. For a review of the history and development of D^3PM , see refs. 3.

The goal of this chapter is to highlight selected key D^3PMs in a practical manner. The focus is on model capabilities and data requirements, aiming to inform the reader's D^3PM analysis strategy based on the desired disease insight(s) and the data available. Figure 1 places selected D^3PMs on a capability×data quadrant

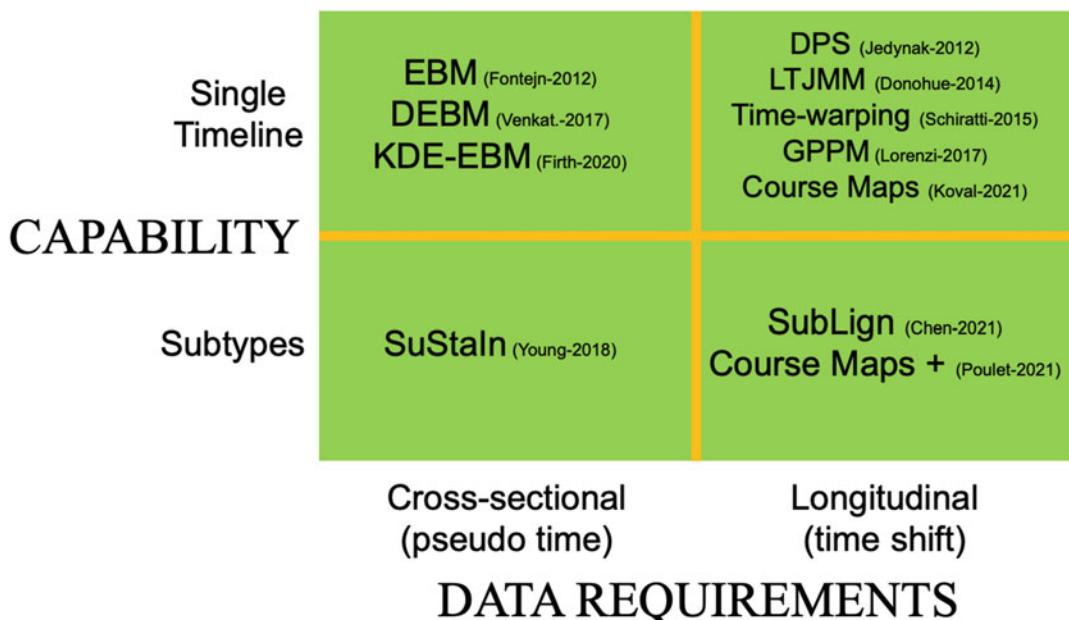


Fig. 1 Quadrant matrix. D^3PMs all estimate a disease timeline, with some capable of estimating multiple subtype timelines, using either cross-sectional data (pseudo-time) or longitudinal data (time-shift). Abbreviations: EBM, event-based model; DEBM, discriminative EBM; KDE-EBM, kernel density estimation EBM; DPS, disease progression score; LTJMM, latent-time joint mixed model; GPPM, Gaussian process progression model; SuStain, subtype and stage inference; SubLign, subtyping alignment

matrix: single timeline estimation vs subtyping, and cross-sectional vs longitudinal data availability. Table A.1 lists more methodological papers relevant to D³PM, with model innovations grouped by the original paper for that method.

The chapter is organized as follows. It starts with a brief discussion of data preprocessing considerations in Subheading 2—an important step in medical data analysis. The treatment of D³PMs is separated into models for cross-sectional data (Subheading 3) and models for longitudinal data (Subheading 4), each split into approaches that estimate a single timeline of disease progression and those capable of estimating multiple timelines within a dataset (subtyping). Subheading 5 concludes.

For a detailed timeline of D³PM development including taxonomy and pedigree of key models, see Appendix.

2 Data Preprocessing

This section briefly touches on two common preprocessing steps before fitting a D³PM to data from a progressive condition such as an irreversible chronic disease: controlling for confounding variables, and handling missing data. We refer to input features as biomarkers and use “covariate” and “confounder” interchangeably. Missing data can refer to irregular/variable visits across individuals, or missing biomarker data due to one or more measurements not being performed for some reason. This section deals with the latter, since longitudinal models can typically handle irregular visits.

Controlling for confounding variables is an important element of any D³PM analysis. This helps to prevent the D³PM from learning non-disease-related patterns such as due to confounding covariates. Confounders can be included as covariates in certain models—to account for that source of variation alongside other variables of interest. Another approach, often used for continuous-valued confounders, is to “regress out” this source of variation prior to fitting a model—to remove non-disease-related signal in the data. This process involves training regression models on data from control participants (who are not expected to develop the disease being studied) and then removing the relevant trends from all data. This method can also be applied to categorical risk factors (discrete variables). The canonical example of a potentially confounding variable in neurodegenerative diseases of the brain is age—a key risk factor in many chronic diseases. Removing normal aging signal is often phrased as “adjusting for” or “controlling for” age.

Handling missing data is an active area of research with a considerable body of literature. Broadly speaking, there are two strategies. The easiest is to exclude participants having any missing biomarker (or covariate) data, but this can considerably reduce the

sample size of data available for D³PM analysis. The second approach is to impute the missing data, e.g., using group mean values. Imputation can be explicit or implicit. An example of implicit imputation is in Bayesian models that map data to probabilities and then deal with missing data probabilistically such as in the event-based model [4] where $P(event|x)=0.5$ represents maximal uncertainty such as when a measurement x is missing.

3 Models for Cross-Sectional Data

Box 1: Models for Cross-Sectional Data

- Pro: Data-economical.
Require cross-sectional data only.
- Con: Limited forecasting utility.
Forecasting requires augmentation with longitudinal data.
- Key application(s): assessing disease severity from a single visit, e.g., economical stratification for clinical research/trials.

3.1 Single Timeline Estimation Using Cross-Sectional Data

3.1.1 Event-Based Model

There is only one framework for estimating disease timelines from cross-sectional data: event-based modeling.

The event-based model (EBM) emerged in 2011 [5, 6]. The concept is simple: in a progressive disease, biomarker measurements only ever get worse, i.e., become increasingly and irreversibly abnormal. Thus, among a cohort of individuals at different stages of a single progressive disease, the cumulative sequence of biomarker abnormality events can be inferred from only a single visit per individual. This requires making a few assumptions: measurements from individuals are independent and represent samples from a single sequence of cumulative abnormality, i.e., a single timeline of disease progression. Such assumptions are commonplace in many statistical analyses of disease progression and are reasonable approximations to make when analyzing data from research studies that typically have strict inclusion and exclusion criteria to focus on a single condition of interest. Unsurprisingly, the event-based model has proven to be extremely powerful, producing insight into many neurodegenerative diseases: sporadic Alzheimer's disease [7–10], familial Alzheimer's disease [6, 11], Huntington's disease [6, 12], Parkinson's disease [13], and others [14, 15].

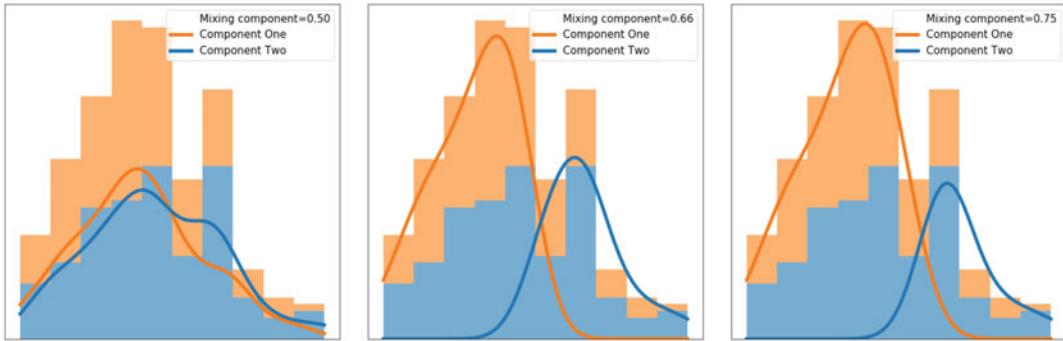


Fig. 2 Event-based models fit a mixture model to map biomarker values to abnormality probabilities. Left to right shows the convergence of a kernel density estimate (KDE) mixture model. From Firth et al. [9] (CC BY 4.0)

EBM Fitting

The first step in fitting an event-based model maps biomarker values to abnormality values, similar to the hypothetical curves of biomarker abnormality proposed in 2010 [1, 2]. The EBM does this probabilistically, using bivariate mixture modeling where individuals can be labeled either as pre-event/normal or post-event/abnormal to allow for (later) events that are yet to occur in patients, and similarly for the possibility of (earlier) events to have occurred in asymptomatic individuals. Various distributions have been proposed for this mixture modeling: combinations of uniform [5, 6], Gaussian [5–7], and kernel density estimate (KDE) distributions [9]. This is visualized in Fig. 2.

The second step in fitting an EBM over N events is to search the space of $N!$ possible sequences S to reveal the most likely sequence (see refs. 6, 7, 9 for mathematical details). For small $N \leq 10$, it can be computationally feasible to perform an exhaustive search over all possible $N!$ sequences to find the maximum likelihood/a posteriori solution. The EBM uses a combination of multiply-initialized gradient ascent, followed by MCMC sampling to estimate uncertainty in the sequence. This results in a model posterior that is a collection of samples from the posterior probability density for each biomarker as a function of sequence position. This is presented as a positional variance diagram [6], such as in Fig. 3.

For further information and to try out EBM tutorials, the reader is directed to the open-source `kde_ebm` package (github.com/ucl-pond/kde_ebm) and disease-progression-modelling.github.io.

3.1.2 Discriminative Event-Based Model

The discriminative event-based model (DEBM) was proposed in 2017 by Venkatraghavan et al. [16]. Whereas the EBM treats data from individuals as observations of a single group-level disease cascade (sequence), the DEBM estimates individual-level sequences and combines them into a group-level description of

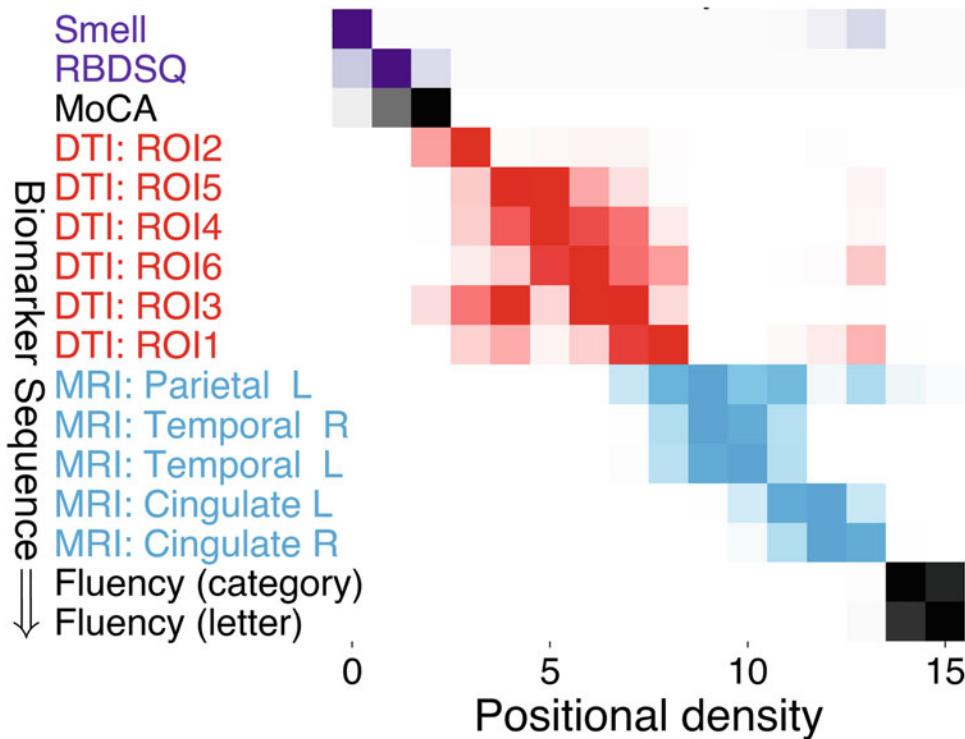


Fig. 3 The event-based model posterior is a positional variance diagram showing uncertainty (left-to-right) in the maximum likelihood sequence (top-to-bottom). Parkinson’s disease model from Oxtoby et al. [13] (CC BY 4.0)

disease progression. This is done using a Mallow’s model, which is the ranking/sequencing equivalent of a univariate Gaussian distribution—including estimation of a mean sequence and variance in this mean. Both EBM and DEBM estimate group-level biomarker abnormality using mixture modeling and both approaches directly estimate uncertainty in the sequence.

Additionally, Venkatraghavan et al. [16, 17] also introduced a pseudo-temporal “disease time” that converts the DEBM posterior into a continuous measure of disease severity.

DEBM Fitting

As with the EBM, DEBM model fitting starts with mixture modeling (see Subheading 3.1.1). Next, a sequence is estimated for each individual by ranking the abnormality probabilities in descending order. A group-level mean sequence (with variance) is estimated by fitting the individual sequences to a Mallow’s model. For details, see refs. 16, 17 and subsequent innovations to the DEBM. Notably, DEBM is often quicker to fit than EBM, which makes it appealing for high-dimensional extensions, e.g., aiming to estimate voxel-wise atrophy signatures from cross-sectional brain imaging data.

For further information and to try it out, the reader is directed to the open-source `pyebm` package (<https://github.com/88vikram/pyebm>).

3.2 Subtyping Using Cross-Sectional Data

Box 2: Subtyping Models

- Pro: Uncovering heterogeneity without conflating severity with subtype.

Evidence suggests that disease subtypes exist.

- Con: Overly simplistic.

Current models ignore comorbidity.

Augmenting the event-based model concept with unsupervised machine learning, subtype and stage inference (SuStaIn), was introduced by Young et al. [18]. This marriage of clustering to disease progression modeling has proven very powerful and popular, with high-impact results appearing in prominent journals for multiple brain diseases [19–21], chronic lung disease [22], and knee osteoarthritis [23]. SuStaIn’s popularity is perhaps unsurprising given that it was the first method capable of unraveling spatiotemporal heterogeneity (pathological severity across an organ) from phenotypic heterogeneity (disease subtypes) in progressive conditions using only cross-sectional data.

Figure 4 (adapted from [18]) shows the concept behind SuStaIn. SuStaIn iteratively solves the clustering problem from 1 to N_S^{\max} subtypes. The N_S model is fitted by splitting each of the $N_S - 1$ subtypes into two clusters and then solving the N_S -cluster problem, which produces $N_S - 1$ candidate N_S -cluster models, from which the maximum likelihood model is chosen, and then the algorithm continues to $N_S + 1$ and so on.

Young et al. [18] also introduced the z-score event progression model that breaks down individual biomarker events into piecewise linear transitions between z-scores of interest. This removes the need for mixture modeling (such as in event-based modeling) and enables inference to be performed at subthreshold biomarker values.

SuStaIn Fitting

For the user, a SuStaIn analysis is very similar to an event-based model analysis. For further information, the reader is directed to the open-source `pySuStaIn` package [24] (<https://github.com/ucl-pond/pySuStaIn>), which includes tutorials. As well as the z-score progression model, `pySuStaIn` includes the various event-based models (see Subheading 3.1), and the more recent scored-events model for ordinal data [25] such as visual ratings of medical images.

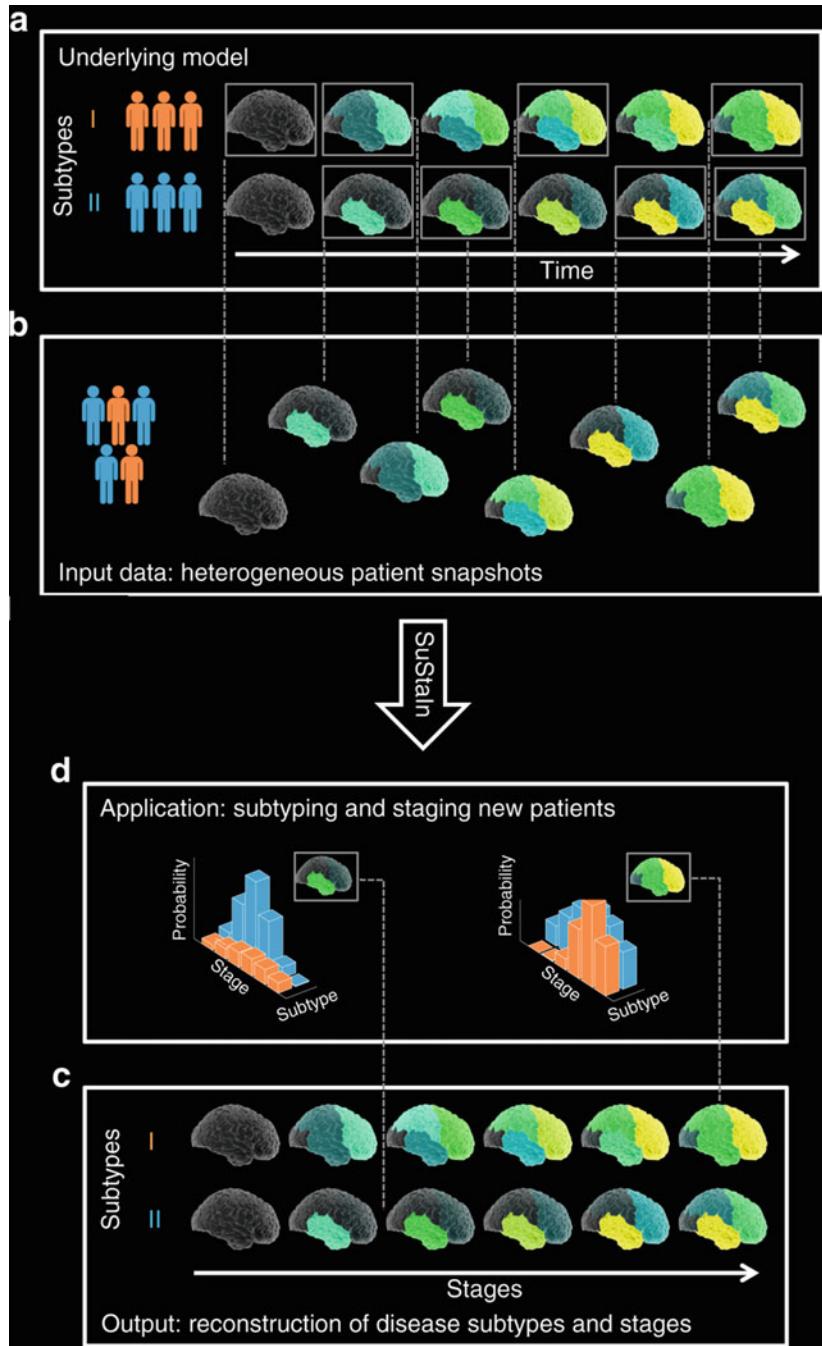


Fig. 4 The concept of subtype and stage inference (SuStain). Reproduced from Young et al. [18] (CC BY 4.0)

4 Models for Longitudinal Data

Box 3: Models for Longitudinal Data

- Pro: Good forecasting utility.
High temporal precision allows individualized forecasting.
- Con: Data-heavy.
Require longitudinal data (multiple visits, years). Can be slow to fit.
- Key application(s): assessing speed of disease progression and assessing individual variability.

The availability of longitudinal data has fueled development of more sophisticated D³PMs, inspired by mixed models. Mixed (effect) modeling is the workhorse of longitudinal statistical analysis against a known timeline, e.g., age. Mixed models provide a hierarchical description of individual-level variation (random effects) about group-level trends (fixed effects), hence the common parlance “mixed-effects” models. Many of the D³PMs for longitudinal data discussed below are in fact mixed models with an additional latent-time parameter that characterizes the disease timeline. Similar approaches in various fields are known as “self-modeling regression” or “latent-time” models. We focus on parametric models, but also mention nonparametric models, and an emerging hybrid discrete-continuous model.

4.1 Single Timeline Estimation Using Longitudinal Data

There are both parametric and nonparametric approaches to estimating disease timelines from longitudinal data. The common goal is to “stitch together” a full disease timeline (decades long) out of relatively short samples from individuals (a few years each) covering a range of severity in symptoms and biomarker abnormality. Some of the earliest work emerged from the medical image registration community, where “warping” images to a common template is one of the first steps in group analyses [26].

Broadly speaking, there are two categories of D³PMs for longitudinal data: time-shifting models and differential equation models. Time-shifting models translate/deform the individual data, metaphorically stitching them together into a quantitative template of disease progression. Differential equation models estimate a statistical model of biomarker dynamics in phase-plane space (position vs velocity), which is subsequently inverted to produce biomarker trajectories.

4.1.1 Explicit Models for Longitudinal Data: Latent-Time Models

Jedynak et al. [27] introduced the disease progression score (DPS) model in 2012, which aligns biomarker data from individuals to a group template model using a linear transformation of age into a disease progression score $s_i = \alpha_i \text{age} + \beta_i$. Individuals have their own rate of progression α_i (constant over the short observation time) and disease onset β_i . Group-level biomarker dynamics are modeled as sigmoid (“S”) curves. A Bayesian extension of the DPS approach (BPS) appeared in 2019 [28]. Code for both the DPS and BPS was released publicly: <https://www.nitrc.org/projects/progscore>; <https://hub.docker.com/r/bilgelm/bayesian-ps-adni/>.

Donohue et al. [29] introduced a self-modeling regression approach similar to the DPS model in 2014. It was later generalized into the more flexible latent-time joint mixed (effects) model (LTJMM) [30], which can include covariates as fixed effects and is a flexible Bayesian framework for inference. The LTJMM software was released publicly: <https://bitbucket.org/mdonohue/lwjglm>.

A nonparametric latent-time mixed model appeared in 2017: the Gaussian process progression model (GPPM) of Lorenzi et al. [31]. This is a flexible Bayesian approach akin to (parametric) self-modeling regression that doesn’t impose a parametric form for biomarker trajectories. More recent work supplemented the GPPM with a dynamical systems model of molecular pathology spread through the brain [32] that can regularize the GPPM fit to produce a more accurate disease timeline reconstruction that also provides insight into neurodegenerative disease mechanisms (which is a topic that could be a standalone chapter of this book). The GPPM and GPPM-DS model source code was released publicly via gitlab.inria.fr/epione and tutorials are available at disease-progression-model.github.io.

In 2015, Schiratti et al. [33–35] introduced a general framework for estimating spatiotemporal trajectories for any type of manifold-valued data. The framework is based on Riemannian geometry and a mixed-effects model with time reparametrization. It was subsequently extended by Koval et al. [36] to form the disease course mapping approach (available in the `leaspy` software package). Disease course mapping combines time warping (of age) and inter-biomarker spacing translation. Time warping changes disease progression dynamics—time shift/onset and acceleration/progression speed—but not the trajectory. Inter-biomarker spacings shift an individual’s trajectory to account for individual differences in the timing and ordering of biomarker trajectories.

Figures 5 and 6 show example outputs of these models when trained on data from older people at risk of Alzheimer’s disease, including those with diagnosed mild cognitive impairment and dementia due to probable Alzheimer’s disease.

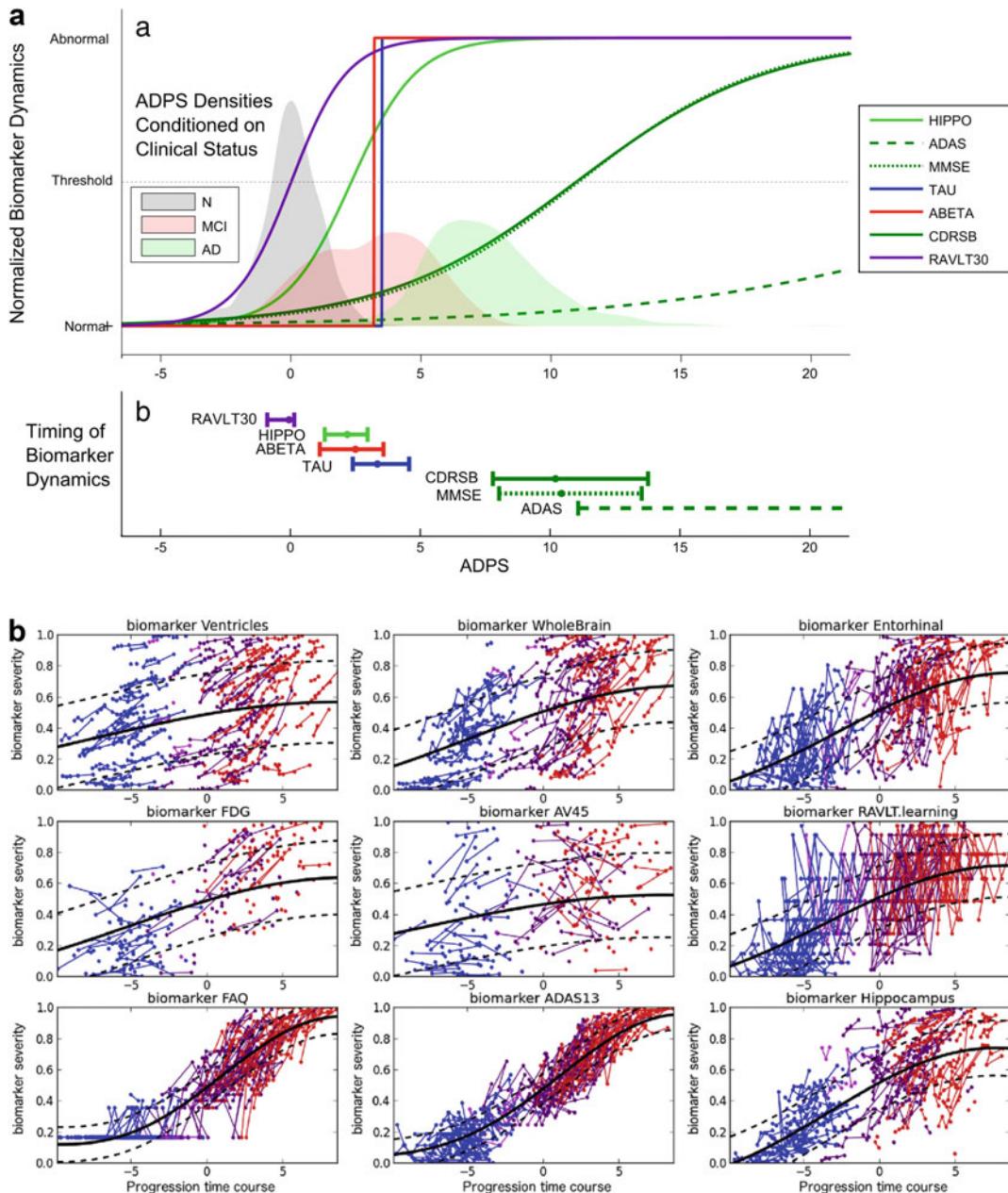


Fig. 5 Two examples of D³PMs fit to longitudinal data: disease progression score [27] and Gaussian process progression model [31]. **(a)** Alzheimer's disease progression score (2012) [27]. Reprinted from *NeuroImage*, Vol 63, Jedynak et al., A computational neurodegenerative disease progression score: Method and results with the Alzheimer's Disease Neuroimaging Initiative cohort, 1478–1486, ©(2012), with permission from Elsevier. **(b)** Gaussian process progression model (2017) [31]. Reprinted from *NeuroImage*, Vol 190, Lorenzi et al., Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease, 56–68, ©(2019), with permission from Elsevier

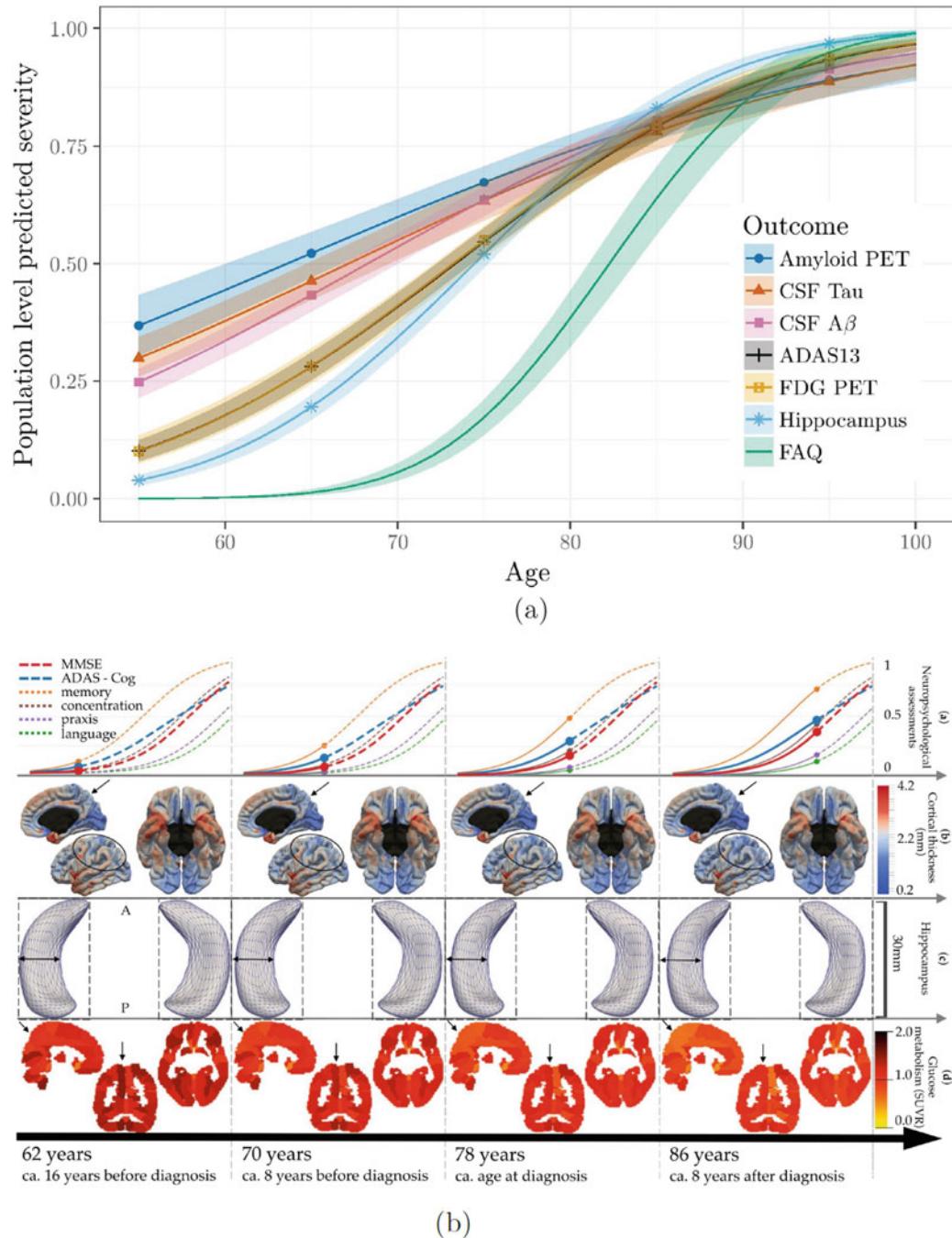


Fig. 6 Two additional examples of D³PMs fit to longitudinal data: latent-time joint mixed model [30] and disease course mapping [36]. **(a)** Latent-time joint mixed model (2017) [30]. From [37] (CC BY 4.0). **(b)** Alzheimer's disease course map (2021) [36] (CC BY 4.0)

Fitting Longitudinal Latent-Time Models

Fitting D³PMs for longitudinal data is more complex than for cross-sectional data, and the software packages discussed above each expect the data in slightly different formats. One thing they have in common is that renormalization (e.g., min-max or z-score) and reorientation (e.g., to be increasing) is required to put biomarkers on a common scale and direction. In some cases, such pre-processing is necessary to ensure/accelerate model convergence. For example, the LTJMM used a quantile transformation followed by inverse Gaussian quantile function to put all biomarkers on a Gaussian scale. For further detailed discussion, including model identifiability, we refer the reader to the original publications cited above and the didactic resources at disease-progression-modelling.github.io.

4.1.2 Implicit Models for Longitudinal Data: Differential Equation Models

Parametric differential equation D³PMs emerged between 2011 and 2014 [38–41], receiving a more formal treatment in 2017 [42]. In a hat-tip to physics, these have also been dubbed “phase-plane” models, which aids in their understanding as a model of velocity (biomarker progression rate) as a function of position (biomarker value). Model fitting is a two-step process whereby the long-time biomarker trajectory is estimated by integrating the phase-plane model estimated on observed data.

A nonparametric differential equation D³PM using Gaussian processes (GP-DEM) was introduced in 2018 [11]. This added flexibility to the preceding parametric approaches and produced state-of-the-art results in predicting symptom onset in familial Alzheimer’s disease.

Fitting Differential Equation Models

The concept is shown in Fig. 7: differential equation model fitting is a three-step process. First, estimate a single value per individual of biomarker “velocity” and “position,” and then estimate a group-level differential equation model of velocity y as a function of position x , which is integrated/inverted to produce a biomarker trajectory $x(t)$. For example, linear regression can produce estimates of position (e.g., intercept) and velocity (e.g., gradient). Differential equation models can be univariate or multivariate and can include covariates explicitly.

4.1.3 Hybrid Discrete-Continuous Models

Recent work introduced the temporal EBM (TEBM) [43, 44], which augments event-based modeling with hidden Markov modeling to produce a hybrid discrete-continuous D³PM. This is a halfway house between discrete models (great for medical decision making) and continuous models (great for detailed understanding of disease progression). Trained on data from ADNI, the TEBM revealed the full timeline of the pathophysiological cascade of Alzheimer’s disease, as shown in Fig. 8.

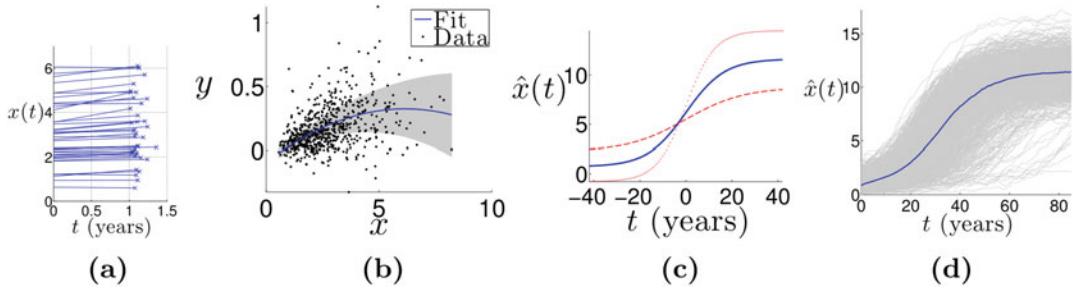


Fig. 7 Differential equation models, or phase-plane models, for biomarker dynamics involve a three-step process: estimate individual-level position and velocity; fit a group-level model of velocity y vs position x ; and integrate to produce a trajectory $x(t)$. Reprinted by permission from Springer Nature: Oxtoby, N.P. et al., Learning Imaging Biomarker Trajectories from Noisy Alzheimer's Disease Data Using a Bayesian Multilevel Model. In: Cardoso, M.J., Simpson, I., Arbel, T., Precup, D., Ribbens, A. (eds) Bayesian and grAphical Models for Biomedical Imaging. Lecture Notes in Computer Science, vol 8677, pp. 85–94 ©(2014) [41]. (a) Data. (b) Differential fit. (c) Est. trajectory. (d) Stochastic model

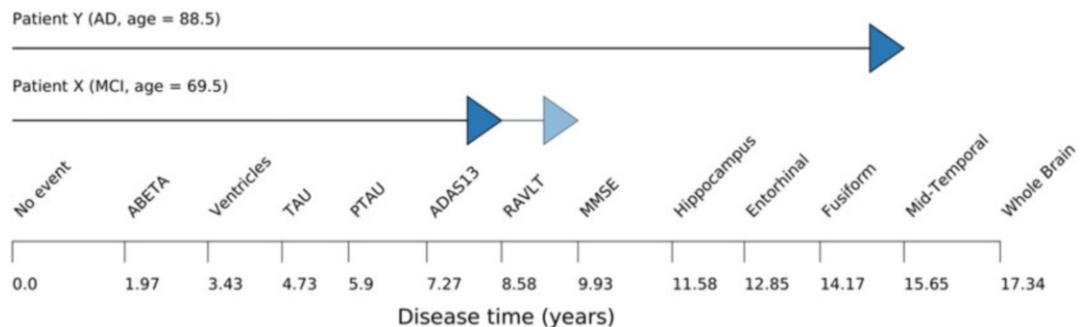


Fig. 8 Alzheimer's disease sequence and timeline estimated by a hybrid discrete-continuous D³PM: the temporal event-based model [43, 44]. Permission to reuse was kindly granted by the authors of [43]

4.2 Subtyping Using Longitudinal Data

Clustering longitudinal data without a well-defined time axis can be extremely difficult. Jointly estimating latent time for multiple trajectories is an identifiability challenge, i.e., multiple parameter combinations can explain the same data. This is particularly challenging when observations span a relatively small fraction of the full disease timeline, as in age-related neurodegenerative diseases.

Chen et al. [45] introduced SubLign for subtyping and aligning longitudinal disease data. The authors frame the challenge eloquently as having misaligned, interval-censored data: left censoring from patients being observed only after disease onset and right censoring from patient dropout in more severe disease. SubLign combines a deep generative model (based on a recurrent neural network [46]) for learning individual latent time-shifts and parametric biomarker trajectories using a variational approach, followed by k-means clustering. It was applied to data from a Parkinson's disease cohort to recover some known clinical phenotypes in new detail.

Poulet and Durrleman [47] recently added mixture-model clustering to the nonlinear mixed model approach of disease course mapping [36]. The framework jointly estimates model parameters and subtypes using a modification of the expectation-maximization algorithm. In simulated data experiments, their approach outperforms a naive baseline. Experiments on real data in Alzheimer’s disease distinguished rapid from slow clinical progression, with minimal differences in biomarker trajectories.

5 Conclusion

Twenty-first century medicine faces many challenges due to aging populations worldwide, including increasing socioeconomic burden from age-related brain disorders like Alzheimer’s disease. Many failed clinical trials fueled intense debate in neurology in the first decade of this century, culminating in the prominent hypothesis of Alzheimer’s disease progression as a pathophysiological cascade of dynamic biomarker events. This inspired the emergence of data-driven disease progression modeling (D^3PM) from the computer science community during the second decade of the twenty-first century—an explosion of quantitative models for neurodegenerative disease progression enabling numerous high-impact insights across multiple brain disorders. The community continues to build and share open-source code (see Box 4) and run machine learning challenges [48–50]. What will the third decade of the twenty-first century bring for this exciting subset of machine learning for brain disorders?

Acknowledgements

The author is a UKRI Future Leaders Fellow (MR/S03546X/1) and acknowledges conversations and collaborations with colleagues from the UCL POND group (<http://pond.cs.ucl.ac.uk>), the Euro-POND consortium (<http://europond.eu>), the E-DADS consortium (<https://e-dads.github.io>; MR/T046422/1; EU JPND), and the open Disease Progression Modeling Initiative (<https://disease-progression-modelling.github.io>). This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No. 666992.

Appendix

A taxonomy and pedigree of key D³PM papers is given in Table A.1. Box 4 contains links to open-source code for D³PMs.

Table A.1

A taxonomy and pedigree of D³PM papers. *Asterisks denote models for cross-sectional data

Reference (first author only)	Description
Ashford, Curren. Psych. Rep. (2001) [51]	Differential equation
Gomeni, Alz. Dem. (2011) [52]	Differential equation
Sabuncu, Arch. Neurol. (2011) [38]	Differential equation
Samtani, J. Clin. Pharmacol. (2012) [39]	Differential equation
Jedynak, NeuroImage (2012) [27]	Progression score (linear)
⇒ Bilgel, IPMI (2015) [53]	
⇒ Bilgel, NIMG (2016) [54]	Latent time mixed effects
⇒ Bilgel, Alz. Dem. DADM (2019) [28]	Bayesian
*Fonteijn, IPMI (2011) [5]; Fonteijn, NIMG (2012) [6]	Cumulative events
⇒ *Young, Brain (2014) [7]	Robust for sporadic disease
⇒ *Venkatraghavan, IPMI (2017) [16]; Venkatraghavan, NIMG (2019) [17]	Individual-level
⇒ *Young, Nat Commun (2018) [18]	+ Subtyping, + Z-score model
⇒ *Young, Frontiers (2021) [55]	+ Scored events model
⇒ *Firth, Alz Dem (2020) [9]	+ Nonparametric events
⇒ Wijeratne, ML4H2020, IPMI (2021) [43, 44]	+ Transition times
Villemagne, Lancet Neurol (2013) [40]	Differential equation
⇒ Budgeon, Stat. in Med. (2017) [42]	Formalism
Durrleman, Int. J. Comput. Vis. (2013) [56]	Time warping
⇒ Schiratti, NeurIPS (2015) [33]; IPMI (2015) [34]; JMLR (2017) [35]	
⇒ Koval, Sci Rep (2021) [36]	Latent-time mixed effects
⇒ Poulet, IPMI (2021) [47]	+Subtyping
Donohue, Alz Dem (2014) [29]	Latent-time fixed effects
⇒ Li, Stat Meth Med Res (2017) [30]	Latent-time mixed effects
Oxtoby, MICCAI (2014) [57]	Differential equation
⇒ Oxtoby, Brain (2018) [11]	+Nonparametric
Guerrero, NeuroImage (2016) [58]	Instantiated mixed effects

(continued)

Table A.1
(continued)

Reference (first author only)	Description
Leoutsakos, JPAD (2016) [59]	Item response theory
Lorenzi, NeuroImage (2017) [31]	Nonparametric latent time
⇒ Garbarino, IPMI (2019) [60]	+Differential equation
⇒ Garbarino, NeuroImage (2021) [32]	Formalism
Marinescu, NeuroImage (2019) [61]	Spatial clustering (c.f. Schiratti/Bilgel)
Petrella, Comp. Math. Meth. Med. (2019) [62]	Differential equation
Abi Nader, Brain Commun. (2021) [63]	Differential equation
Chen, AAAI (2022) [45]	Subtyping

Box 4: Example Open-Source D³PM Code

- D³PM tutorials:
<https://disease-progression-modelling.github.io>
- EuroPOND Software Toolbox:
<https://europond.github.io/europond-software>
- KDE EBM:
https://ucl-pond.github.io/kde_ebm
- pyEBM:
<https://github.com/88vikram/pyebm>
- leaspy:
<https://gitlab.com/icm-institute/aramislab/leaspy>
- LTJMM:
<https://bitbucket.org/mdonohue/ltjmm>
<https://github.com/mcdonohue/rstanarm>
- DPS:
[source code](#); [docker image](#)
- pySuStaIn:
<https://ucl-pond.github.io/pySuStaIn>
- TADPOLE-SHARE (from TADPOLE Challenge [48, 49]):
<https://github.com/tadpole-share/tadpole-algorithms>

References

1. Jack CR, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, Petersen RC, Trojanowski JQ (2010) Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol* 9(1):119–128. [https://doi.org/10.1016/S1474-4422\(09\)70299-6](https://doi.org/10.1016/S1474-4422(09)70299-6)
2. Aisen PS, Petersen RC, Donohue MC, Gamst A, Raman R, Thomas RG, Walter S, Trojanowski JQ, Shaw LM, Beckett LA, Jr CRJ, Jagust W, Toga AW, Saykin AJ, Morris JC, Green RC, Weiner MW (2010) Clinical core of the Alzheimer's disease neuroimaging initiative: progress and plans. *Alzheimer's Dementia* 6(3):239–246. <https://doi.org/10.1016/j.jalz.2010.03.006>
3. Oxtoby NP, Alexander DC, EuroPOND Consortium (2017) Imaging plus X: multimodal models of neurodegenerative disease. *Curr Opin Neurol* 30(4). <https://doi.org/10.1097/WCO.0000000000000460>
4. Young AL, Oxtoby NP, Ourselin S, Schott JM, Alexander DC (2015) A simulation system for biomarker evolution in neurodegenerative disease. *Med Image Anal* 26(1):47–56. <https://doi.org/10.1016/j.media.2015.07.004>
5. Fonteijn H, Clarkson M, Modat M, Barnes J, Lehmann M, Ourselin S, Fox N, Alexander D (2011) An event-based disease progression model and its application to familial Alzheimer's disease. In: Székely G, Hahn H (eds) Information processing in medical imaging. Lecture notes in computer science, vol 6801. Springer, Berlin/Heidelberg, pp 748–759. https://doi.org/10.1007/978-3-642-22092-0_61
6. Fonteijn HM, Modat M, Clarkson MJ, Barnes J, Lehmann M, Hobbs NZ, Scahill RI, Tabrizi SJ, Ourselin S, Fox NC, Alexander DC (2012) An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImage* 60(3):1880–1889. <https://doi.org/10.1016/j.neuroimage.2012.01.062>
7. Young AL, Oxtoby NP, Daga P, Cash DM, Fox NC, Ourselin S, Schott JM, Alexander DC (2014) A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* 137(9):2564–2577. <https://doi.org/10.1093/brain/awu176>
8. Oxtoby NP, Garbarino S, Firth NC, Warren JD, Schott JM, Alexander DC, FtADNI (2017) Data-driven sequence of changes to anatomical brain connectivity in sporadic Alzheimer's disease. *Front Neurol* 8:580. <https://doi.org/10.3389/fneur.2017.00580>
9. Firth NC, Primativo S, Brotherhood E, Young AL, Yong KX, Crutch SJ, Alexander DC, Oxtoby NP (2020) Sequences of cognitive decline in typical Alzheimer's disease and posterior cortical atrophy estimated using a novel event-based model of disease progression. *Alzheimer's Dementia* 16(7):965–973. <https://doi.org/10.1002/alz.12083>
10. Janelidze S, Berron D, Smith R, Strandberg O, Proctor NK, Dage JL, Stomrud E, Palmqvist S, Mattsson-Carlsgren N, Hansson O (2021) Associations of plasma phospho-Tau217 levels with tau positron emission tomography in early Alzheimer disease. *JAMA Neurol* 78(2):149–156. <https://doi.org/10.1001/jamaneurol.2020.4201>
11. Oxtoby NP, Young AL, Cash DM, Benzinger TLS, Fagan AM, Morris JC, Bateman RJ, Fox NC, Schott JM, Alexander DC (2018) Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain* 141(5):1529–1544. <https://doi.org/10.1093/brain/awy050>
12. Wijeratne PA, Young AL, Oxtoby NP, Marinescu RV, Firth NC, Johnson EB, Mohan A, Sampaio C, Scahill RI, Tabrizi SJ, Alexander DC (2018) An image-based model of brain volume biomarker changes in Huntington's disease. *Ann Clin Transl Neurol* 5(5):570–582. <https://doi.org/10.1002/acn3.558>
13. Oxtoby NP, Leyland LA, Aksman LM, Thomas GEC, Bunting EL, Wijeratne PA, Young AL, Zarkali A, Tan MMX, Bremner FD, Keane PA, Morris HR, Schrag AE, Alexander DC, Weil RS (2021) Sequence of clinical and neurodegeneration events in Parkinson's disease progression. *Brain* 144(3):975–988. <https://doi.org/10.1093/brain/awaa461>
14. Eshaghi A, Marinescu RV, Young AL, Firth NC, Prados F, Jorge Cardoso M, Tur C, De Angelis F, Cawley N, Brownlee WJ, De Stefano N, Laura Stromillo M, Battaglini M, Ruggieri S, Gasperini C, Filippi M, Rocca MA, Rovira A, Sastre-Garriga J, Geurts JJG, Vrenken H, Wottschel V, Leurs CE, Uitdehaag B, Pirpamer L, Enzinger C, Ourselin S, Gandini Wheeler-Kingshott CA, Chard D, Thompson AJ, Barkhof F, Alexander DC, Ciccarelli O (2018) Progression of regional grey matter atrophy in multiple sclerosis. *Brain* 141(6):1665–1677. <https://doi.org/10.1093/brain/awy088>
15. Firth NC, Startin CM, Hithersay R, Hamburg S, Wijeratne PA, Mok KY, Hardy J, Alexander DC, Consortium TL, Strydom A

- (2018) Aging related cognitive changes associated with Alzheimer's disease in Down syndrome. *Ann Clin Transl Neurol* 5(6):741–751. <https://doi.org/10.1002/acn3.571>
16. Venkatraghavan V, Bron EE, Niessen WJ, Klein S (2017) A discriminative event based model for Alzheimer's disease progression modeling. In: Niethammer M, Styner M, Aylward S, Zhu H, Oguz I, Yap PT, Shen D (eds) *Information processing in medical imaging. Lecture notes in computer science*. Springer, Cham, pp 121–133. https://doi.org/10.1007/978-3-319-59050-9_10
 17. Venkatraghavan V, Bron EE, Niessen WJ, Klein S (2019) Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. *NeuroImage* 186: 518–532. <https://doi.org/10.1016/j.neuroimage.2018.11.024>
 18. Young AL et al (2018) Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nat Commun* 9(1):4273. <https://doi.org/10.1038/s41467-018-05892-0>
 19. Vogel JW, Young AL, Oxtoby NP, Smith R, Ossenkoppele R, Strandberg OT, La Joie R, Aksman LM, Grothe MJ, Iturria-Medina Y, the Alzheimer's Disease Neuroimaging Initiative, Pontecorvo MJ, Devous MD, Rabinovici GD, Alexander DC, Lyoo CH, Evans AC, Hansson O (2021) Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med.* <https://doi.org/10.1038/s41591-021-01309-6>
 20. Eshaghi A, Young AL, Wijeratne PA, Prados F, Arnold DL, Narayanan S, Guttmann CRG, Barkhof F, Alexander DC, Thompson AJ, Chard D, Ciccarelli O (2021) Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. *Nat Commun* 12(1):2078. <https://doi.org/10.1038/s4167-021-22265-2>
 21. Collij LE, Salvadó G, Wottschel V, Mastenbroek SE, Schoenmakers P, Heeman F, Aksman L, Wink AM, Berckel BNM, Flier WMvd, Scheltens P, Visser PJ, Barkhof F, Haller S, Gispert JD, Alves IL, for the Alzheimer's Disease Neuroimaging Initiative; for the ALFA Study (2022) Spatial-temporal patterns of β -amyloid accumulation: a subtype and stage inference model analysis. *Neurology* 98(17): e1692–e1703. <https://doi.org/10.1212/WNL.0000000000200148>
 22. Young AL, Bragman FJS, Rangelov B, Han MK, Galbán CJ, Lynch DA, Hawkes DJ, Alexander DC, Hurst JR (2020) Disease progression modeling in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 201(3):294–302. <https://doi.org/10.1164/rccm.201908-1600OC>
 23. Li M, Lan L, Luo J, Peng L, Li X, Zhou X (2021) Identifying the phenotypic and temporal heterogeneity of knee osteoarthritis: data from the osteoarthritis initiative. *Front Public Health* 9. <https://doi.org/10.3389/fpubh.2021.726140>
 24. Aksman LM, Wijeratne PA, Oxtoby NP, Eshaghi A, Shand C, Altmann A, Alexander DC, Young AL (2021) pySuStaIn: a python implementation of the subtype and stage inference algorithm. *SoftwareX* 16:100811. <https://doi.org/10.1016/j.softx.2021.100811>
 25. Young AL, Vogel JW, Aksman LM, Wijeratne PA, Eshaghi A, Oxtoby NP, Williams SCR, Alexander DC, ftADNI (2021) Ordinal sustain: subtype and stage inference for clinical scores, visual ratings, and other ordinal data. *Front Artif Intell* 4:111. <https://doi.org/10.3389/frai.2021.613261>
 26. Durrleman S, Pennec X, Trouvé A, Gerig G, Ayache N (2009) Spatiotemporal atlas estimation for developmental delay detection in longitudinal datasets. In: Yang GZ, Hawkes D, Rueckert D, Noble A, Taylor C (eds) *Medical image computing and computer-assisted intervention—MICCAI 2009*. Springer, Berlin, pp 297–304. https://doi.org/10.1007/978-3-642-04268-3_37
 27. Jedynak BM, Lang A, Liu B, Katz E, Zhang Y, Wyman BT, Raunig D, Jedynak CP, Caffo B, Prince JL (2012) A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease neuroimaging initiative cohort. *NeuroImage* 63(3): 1478–1486. <https://doi.org/10.1016/j.neuroimage.2012.07.059>
 28. Bilgel M, Jedynak BM, Alzheimer's Disease Neuroimaging Initiative (2019) Predicting time to dementia using a quantitative template of disease progression. *Alzheimer's Dementia: Diagn Assess Dis Monit* 11(1):205–215. <https://doi.org/10.1016/j.jadm.2019.01.005>
 29. Donohue MC, Jacqmin-Gadda H, Goff ML, Thomas RG, Raman R, Gamst AC, Beckett LA, Jack Jr. CR, Weiner MW, Dartigues JF, Aisen PS (2014) Estimating long-term multivariate progression from short-term data. *Alzheimer's Dementia* 10(5, Supplement):S400–S410. <https://doi.org/10.1016/j.jalz.2013.10.003>
 30. Li D, Iddi S, Thompson WK, Donohue MC (2019) Bayesian latent time joint mixed effect

- models for multicohort longitudinal data. *Stat Methods Med Res* 28(3):835–845. <https://doi.org/10.1177/0962280217737566>
31. Lorenzi M, Filippone M, Frisoni GB, Alexander DC, Ourselin S (2019) Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer's disease. *NeuroImage* 190:56–68. <https://doi.org/10.1016/j.neuroimage.2017.08.059>
 32. Garbarino S, Lorenzi M (2021) Investigating hypotheses of neurodegeneration by learning dynamical systems of protein propagation in the brain. *NeuroImage* 235:117980. <https://doi.org/10.1016/j.neuroimage.2021.117980>
 33. Schiratti JB, Allassonnière S, Colliot O, Durrleman S (2015) Learning spatiotemporal trajectories from manifold-valued longitudinal data. In: Advances in neural information processing systems, Curran Associates, vol 28. <https://proceedings.neurips.cc/paper/2015/hash/186a157b2992e7daed3677ce8e9fe40f-Abstract.html>
 34. Schiratti JB, Allassonnière S, Routier A, Colliot O, Durrleman S (2015) A mixed-effects model with time reparametrization for longitudinal univariate manifold-valued data. In: Ourselin S, Alexander DC, Westin CF, Cardoso MJ (eds) Information processing in medical imaging. Lecture notes in computer science. Springer, Cham, pp 564–575. https://doi.org/10.1007/978-3-319-19992-4_44
 35. Schiratti JB, Allassonnière S, Colliot O, Durrleman S (2017) A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. *J Mach Learn Res* 3:1–48
 36. Koval I, Bône A, Louis M, Lartigue T, Bottani S, Marcoux A, Samper-González J, Burgos N, Charlier B, Bertrand A, Epelbaum S, Colliot O, Allassonnière S, Durrleman S (2021) AD course map charts Alzheimer's disease progression. *Sci Rep* 11(1):8020. <https://doi.org/10.1038/s41598-021-87434-1>
 37. Li D, Iddi S, Thompson WK, Donohue MC (2017) Bayesian latent time joint mixed effect models for multicohort longitudinal data. arXiv 1703.10266v2. <https://doi.org/10.48550/arXiv.1703.10266>
 38. Sabuncu M, Desikan R, Sepulcre J, Yeo B, Liu H, Schmansky N, Reuter M, Weiner M, Buckner R, Sperling R (2011) The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Arch. Neurol.* 68(8):1040. <https://doi.org/10.1001/archneurol.2011.167>
 39. Samtani MN, Farnum M, Lobanov V, Yang E, Raghavan N, DiBernardo A, Narayan V, Alzheimer's Disease Neuroimaging Initiative (2012) An improved model for disease progression in patients from the Alzheimer's disease neuroimaging initiative. *J Clin Pharmacol* 52(5):629–644. <https://doi.org/10.1177/0091270011405497>
 40. Villemagne VL, Burnham S, Bourgeat P, Brown B, Ellis KA, Salvador O, Szoek C, Macaulay SL, Martins R, Maruff P, Ames D, Rowe CC, Masters CL (2013) Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. *Lancet Neurol* 12(4): 357–367. [https://doi.org/10.1016/S1474-4422\(13\)70044-9](https://doi.org/10.1016/S1474-4422(13)70044-9)
 41. Oxtoby NP, Young AL, Fox NC, Daga P, Cash DM, Ourselin S, Schott JM, Alexander DC (2014) Learning imaging biomarker trajectories from noisy Alzheimer's disease data using a Bayesian multilevel model. In: Cardoso MJ, Simpson I, Arbel T, Precup D, Ribbens A (eds) Bayesian and graphical models for biomedical imaging. Lecture notes in computer science, vol 8677. Springer, Berlin, pp 85–94. https://doi.org/10.1007/978-3-319-12289-2_8
 42. Budgeon C, Murray K, Turlach B, Baker S, Villemagne V, Burnham S, for the Alzheimer's Disease Neuroimaging Initiative (2017) Constructing longitudinal disease progression curves using sparse, short-term individual data with an application to Alzheimer's disease. *Stat Med* 36(17):2720–2734. <https://doi.org/10.1002/sim.7300>
 43. Wijeratne PA, Alexander DC (2020) Learning transition times in event sequences: the event-based hidden Markov model of disease progression. <https://doi.org/10.48550/arXiv.2011.01023>. Machine Learning for Health (ML4H) 2020
 44. Wijeratne PA, Alexander DC (2021) Learning transition times in event sequences: the temporal event-based model of disease progression. In: Feragen A, Sommer S, Schnabel J, Nielsen M (eds) Information processing in medical imaging. Lecture notes in computer science. Springer, Cham, pp 583–595. https://doi.org/10.1007/978-3-030-78191-0_45
 45. Chen IY, Krishnan RG, Sontag D (2022) Clustering interval-censored time-series for disease phenotyping. In: Proceedings of the AAAI conference on artificial intelligence, vol 36(6), pp 6211–6221. <https://doi.org/10.1609/aaai.v36i6.20570>

46. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088): 533–536. <https://doi.org/10.1038/323533a0>
47. Poulet PE, Durrleman S (2021) Mixture modeling for identifying subtypes in disease course mapping. In: Feragen A, Sommer S, Schnabel J, Nielsen M (eds) *Information processing in medical imaging. Lecture notes in computer science*. Springer, Cham, pp 571–582. https://doi.org/10.1007/978-3-030-78191-0_44
48. Marinescu RV, Oxtoby NP, Young AL, Bron EE, Toga AW, Weiner MW, Barkhof F, Fox NC, Klein S, Alexander DC, EuroPOND Consortium (2018) TADPOLE challenge: prediction of longitudinal evolution in Alzheimer’s disease. <https://doi.org/10.48550/arXiv.1805.03909>
49. Marinescu RV, Oxtoby NP, Young AL, Bron EE, Alexander DC et al (2021) The Alzheimer’s disease prediction of longitudinal evolution (TADPOLE) challenge: results after 1 year follow-up. *Mach Learn Biomed Imaging* 1:1–10. <https://www.melba-journal.org/papers/2021:019.html>
50. Bron EE, Klein S, Reinke A, Papma JM, Maier-Hein L, Alexander DC, Oxtoby NP (2022) Ten years of image analysis and machine learning competitions in dementia. *NeuroImage* 253:119083. <https://doi.org/10.1016/j.neuroimage.2022.119083>
51. Ashford JW, Schmitt FA (2001) Modeling the time-course of Alzheimer dementia. *Current Psychiatry Rep* 3(1):20–28. <https://doi.org/10.1007/s11920-001-0067-1>
52. Gomeni R, Simeoni M, Zwartau-Hind M, Irizarry MC, Austin D, Gold M (2012) Modeling Alzheimer’s disease progression using the disease system analysis approach. *Alzheimer’s Dementia* 8(1):39–50. <https://doi.org/10.1016/j.jalz.2010.12.012>
53. Bilgel M, Jedynak B, Wong DF, Resnick SM, Prince JL (2015) Temporal trajectory and progression score estimation from voxelwise longitudinal imaging measures: application to amyloid imaging. In: Ourselin S, Alexander DC, Westin CF, Cardoso MJ (eds) *Information processing in medical imaging. Lecture notes in computer science*. Springer, Cham, pp 424–436. https://doi.org/10.1007/978-3-319-19992-4_33
54. Bilgel M, Prince JL, Wong DF, Resnick SM, Jedynak BM (2016) A multivariate nonlinear mixed effects model for longitudinal image analysis: application to amyloid imaging. *NeuroImage* 134:658–670. <https://doi.org/10.1016/j.neuroimage.2016.04.001>
55. Young AL, Vogel JW, Aksman LM, Wijeratne PA, Eshaghi A, Oxtoby NP, Williams SCR, Alexander DC, for the Alzheimer’s Disease (2021) Ordinal SuStaIn: subtype and stage inference for clinical scores, visual ratings, and other ordinal data. *Front Artif Intell* 4(613261). <https://doi.org/10.3389/frai.2021.613261>
56. Durrleman S, Pennec X, Trouvé A, Braga J, Gerig G, Ayache N (2013) Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. *Int J Comput Vis* 103(1):22–59. <https://doi.org/10.1007/s11263-012-0592-x>
57. Oxtoby NP, Young AL, Fox NC, Daga P, Cash DM, Ourselin S, Schott JM, Alexander DC (2014) Learning imaging biomarker trajectories from noisy Alzheimer’s disease data using a Bayesian multilevel model. In: Cardoso MJ, Simpson I, Arbel T, Precup D, Ribbens A (eds) *Bayesian and graphical models for biomedical imaging*. Springer, Cham, pp 85–94. https://doi.org/10.1007/978-3-319-12289-2_8
58. Guerrero R, Schmidt-Richberg A, Ledig C, Tong T, Wolz R, Rueckert D (2016) Instantiated mixed effects modeling of Alzheimer’s disease markers. *NeuroImage* 142:113–125. <https://doi.org/10.1016/j.neuroimage.2016.06.049>
59. Leoutsakos JM, Gross AL, Jones RN, Albert MS, Breitner JCS (2016) ‘Alzheimer’s progression score’: development of a biomarker summary outcome for AD prevention trials. *J Prev Alzheimer’s Dis* 3(4):229–235. <https://doi.org/10.14283/jpad.2016.120>
60. Garbarino S, Lorenzi M (2019) Modeling and inference of spatio-temporal protein dynamics across brain networks. In: Chung ACS, Gee JC, Yushkevich PA, Bao S (eds) *Information processing in medical imaging. Lecture notes in computer science*. Springer, Cham, pp 57–69. https://doi.org/10.1007/978-3-030-20351-1_5
61. Marinescu RV, Eshaghi A, Lorenzi M, Young AL, Oxtoby NP, Garbarino S, Crutch SJ, Alexander DC (2019) DIVE: a spatiotemporal progression model of brain pathology in neurodegenerative disorders. *NeuroImage* 192:166–177. <https://doi.org/10.1016/j.neuroimage.2019.02.053>
62. Petrella JR, Hao W, Rao A, Doraiswamy PM (2019) Computational causal modeling of the dynamic biomarker cascade in Alzheimer’s disease. *Comput Math Methods Med* 2019:

- e6216530. <https://doi.org/10.1155/2019/6216530>
63. Abi Nader C, Ayache N, Frisoni GB, Robert P, Lorenzi M, for the Alzheimer's Disease Neuro-imaging Initiative (2021) Simulating the

outcome of amyloid treatments in Alzheimer's disease from imaging and clinical data. *Brain Commun* 3(2):fcab091. <https://doi.org/10.1093/braincomms/fcab091>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made. The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 18

Computational Pathology for Brain Disorders

Gabriel Jiménez and Daniel Racoceanu

Abstract

Noninvasive brain imaging techniques allow understanding the behavior and macro changes in the brain to determine the progress of a disease. However, computational pathology provides a deeper understanding of brain disorders at cellular level, able to consolidate a diagnosis and make the bridge between the medical image and the omics analysis. In traditional histopathology, histology slides are visually inspected, under the microscope, by trained pathologists. This process is time-consuming and labor-intensive; therefore, the emergence of computational pathology has triggered great hope to ease this tedious task and make it more robust. This chapter focuses on understanding the state-of-the-art machine learning techniques used to analyze whole slide images within the context of brain disorders. We present a selective set of remarkable machine learning algorithms providing discriminative approaches and quality results on brain disorders. These methodologies are applied to different tasks, such as monitoring mechanisms contributing to disease progression and patient survival rates, analyzing morphological phenotypes for classification and quantitative assessment of disease, improving clinical care, diagnosing tumor specimens, and intraoperative interpretation. Thanks to the recent progress in machine learning algorithms for high-content image processing, computational pathology marks the rise of a new generation of medical discoveries and clinical protocols, including in brain disorders.

Key words Computational pathology, Digital pathology, Whole slide imaging, Machine learning, Deep learning, Brain disorders

1 Introduction

1.1 What Are We Presenting?

This chapter aims to assist the reader in discovering and understanding state-of-the-art machine learning techniques used to analyze whole slide images (WSI), an essential data type used in computational pathology (CP). We are restricting our review to brain disorders, classified within four generally accepted groups:

- *Brain injuries*: caused by blunt trauma and can damage brain tissue, neurons, and nerves.
- *Brain tumors*: can originate directly from the brain (and be cancerous or benign) or be due to metastasis (cancer elsewhere in the body and spreading to the brain).

- *Neurodegenerative diseases*: the brain and nerves deteriorate over time. We include, here, Alzheimer's disease, Huntington's disease, ALS (amyotrophic lateral sclerosis) or Lou Gehrig's disease, and Parkinson's disease.
- *Mental disorders*: (or mental illness) affect behavior patterns. Depression, anxiety, bipolar disorder, PTSD (post-traumatic stress disorder), and schizophrenia are common diagnoses.

In the last decade, there has been exponential growth in the application of image processing and artificial intelligence (AI) algorithms within digital pathology workflows. The first FDA (US Food and Drug Administration) clearance of digital pathology for diagnosis protocols was as early as 2017,¹ as the emergence of innovative deep learning (DL) technologies have made this possible, with the requested degree of robustness and repeatability.

Ahmed Serag et al. [1] discuss the translation of AI into clinical practice to provide pathologists with new tools to improve diagnostic consistency and reduce errors. In the last five years, the authors reported an increase in academic publications (over 1000 articles reported in PubMed) and over \$100M invested in start-ups building practical AI applications for diagnostics. The three main areas of development are (i) *network architectures* to extract relevant features from WSI for classification or segmentation purposes, (ii) *generative adversarial networks (GANs)* to address some of the issues present in the preparation and acquisition of WSIs, and (iii) *unsupervised learning* to create labeling tools for precise annotations. Regarding data, many top-tier conference competitions have been organized and released annotated datasets to the community; however, very few of them contain brain tissue samples. Those which do are from brain tumor regions obtained during a biopsy, making it harder to study other brain disorder categories which frequently require postmortem data.

In [1], the authors also mention seven key challenges in diagnostic AI in pathology, listed as follows:

- Access to large well-annotated datasets. Most articles on brain disorders use private datasets due to hospital privacy constraints.
- Context switching between workflows refers to a seamless integration of AI into the pathology workflow.
- Algorithms are slow to run as image sizes are in gigapixels' order and require considerable computational memory.

¹ <https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-whole-slide-imaging-system-digital-pathology>.

- Algorithms require configuration, and fully automated approaches with high accuracy are difficult to develop.
- Properly defined protocols are needed for training and evaluation.
- Algorithms are not properly validated due to a lack of open datasets. However, research in data augmentation might help in this regard.
- Introduction of intelligence augmentation to describe computational pathology improvements in diagnostic pathology. AI algorithms work best on well-defined domains rather than in the context of multiple clinicopathological manifestations among a broad range of diseases; however, they provide relevant quantitative insights needed for standardization and diagnosis.

These challenges limit the translation from research to clinical diagnostics. We intend to give the readers some insights into the core problems behind the issues listed by briefly introducing WSI preparation and image acquisition protocols. Besides, we describe the state of the art of the proposed methods.

1.2 Why AI for Brain Disorders?

An important role of CP in brain disorders is related to the study and assessment of brain tumors as they cause significant morbidity and mortality worldwide, and pathology data is often available. In 2022 [2], over 25k adults (14,170 men and 10,880 women) in the United States will have been diagnosed with primary cancerous tumors of the brain and spinal cord. 85% to 90% of all primary central nervous system (CNS) tumors (benign and cancerous) are located in the brain. Worldwide, over 300k people were diagnosed with a primary brain or spinal cord tumor in 2020. This disorder does not distinguish age, as nearly 4.2k children under the age of 15 will have also been diagnosed with brain or CNS tumors in 2022, in the United States.

It is estimated that around one billion people have a mental or substance use disorder [3]. Some other key figures related to mental disorders worldwide are given by [4]. Globally, an estimated 264 million people are affected by depression. Bipolar disorder affects about 45 million people worldwide. Schizophrenia affects 20 million people worldwide, and approximately 50 million have dementia. In Europe, an estimated 10.5 million people have dementia, and this number is expected to increase to 18.7 million in 2050 [5].

In the neurodegenerative disease group, 50 million people worldwide are living with Alzheimer's and other types of dementia [6], Alzheimer's disease being the underlying cause in 70% of people with dementia [5]. Parkinson's disease affects approximately 6.2 million people worldwide [7] and represents the second most common neurodegenerative disorder. As the incidence of

Alzheimer's and Parkinson's diseases rises significantly with age and people's life expectancy has increased, the prevalence of such disorders is set to rise dramatically in the future. For instance, there may be nearly 13 million people with Parkinson's by 2040 [7].

Brain injuries are also the subject of a considerable number of incidents. Every year, around 17 million people suffer a stroke worldwide, with an estimate of one in four persons having a stroke during their lifetime [8]. Besides, stroke is the second cause of death worldwide and the first cause of acquired disability [5].

These disorders also impact American regions, with over 500k deaths reported in 2019, due to neurological conditions. Among the conditions analyzed, the most common ones were Alzheimer's disease, Parkinson's, epilepsy, and multiple sclerosis [9].

In the case of brain tumors, treatment and prognosis require accurate and expedient histological diagnosis of the patient's tissue samples. Trained pathologists visually inspect histology slides, following a time-consuming and labor-intensive procedure. Therefore, the emergence of CP has triggered great hope to ease this tedious task and make it more robust. Clinical workflows in oncology rely on predictive and prognostic molecular biomarkers. However, the growing number of these complex biomarkers increases the cost and the time for decision-making in routine daily practice. Available tumor tissue contains an abundance of clinically relevant information that is currently not fully exploited, often requiring additional diagnostic material. Histopathological images contain rich phenotypic information that can be used to monitor underlying mechanisms contributing to disease progression and patient survival outcomes.

In most other brain diseases, histological images are only acquired postmortem, and this procedure is far from being systematic. Indeed, it depends on the previous agreement of the patient to donate their brain for research purposes. Moreover, as mentioned above, the inspection of such images is complex and tedious, which further explains why it is performed in a minority of cases. Nevertheless, histopathological information is of the utmost importance in understanding the pathophysiology of most neurological disorders, and research progress would be impossible without such images. Finally, there are a few examples, beyond brain tumors, in which a surgical operation leads to an inspection of resected when the patient is alive (this is, for instance, the case of pharmacoresistant epilepsy).

Intraoperative decision-making also relies significantly on histological diagnosis, which is often established when a small specimen is sent for immediate interpretation by a neuropathologist. In poor-resource settings, access to specialists may be limited, which has prompted several groups to develop machine learning (ML) algorithms for automated interpretation. Computerized analysis of digital pathology images offers the potential to improve

clinical care (e.g., automated assistive diagnosis) and catalyze research (e.g., discovering disease subtypes or understanding the pathophysiology of a brain disorder).

1.3 How Do We Present the Information?

In order to understand the potential and limitations of computational pathology algorithms, one needs to understand the basics behind the preparation of tissue samples and the image acquisition protocols followed by scanner manufacturers. Therefore, we have structured the chapter as follows.

Subheading 2 presents an overview of tissue preservation techniques and how they may impact the final whole slide image. Subheading 3 introduces the notion of digital pathology and computational pathology, and its differences. It also develops the image acquisition protocol and describes the pyramidal structure of the WSI and its benefits. In addition, it discusses the possible impact of scanners on image processing algorithms. Subheading 4 describes some of the state-of-the-art algorithms in artificial intelligence and its subcategories (machine learning and deep learning). This section is divided into methods for classifying and segmenting structures in WSI, and techniques that leverage deep learning algorithms to extract meaningful features from the WSI and apply them to a specific clinical application. Finally, Subheading 5 explores new horizons in digital and computational pathology regarding explainability and new microscopic imaging modalities to improve tissue visualization and information retrieval.

2 Understanding Histological Images

We dedicate this section to understanding the process of acquiring histological images. We begin by introducing the two main tissue preservation techniques used in neuroscience studies, i.e., the routine-FFPE (formalin-fixed paraffin-embedded) preparation and the frozen tissue. We describe the process involved in each method and the main limitations for obtaining an appropriate histopathological image for analysis. Finally, we present the main procedures used in anatomopathology, based on such tissue preparations.

2.1 Formalin-Fixed Paraffin-Embedded Tissue

FFPE is a technique used for preserving biopsy specimens for clinical examination, diagnostic, experimental research, and drug development. A correct histological analysis of tissue morphology and biomarker localization in tissue samples will hinge on the ability to achieve high-quality preparation of tissue samples, which usually requires three critical stages: fixation, processing (also known as pre-embedding), and embedding.

Fixation is the process that allows the preservation of the tissue architecture (i.e., its cellular components, extracellular material, and molecular elements). Histotechnologists perform this

procedure right after removing the tissue, in case of surgical pathology, or soon after death, during autopsy. Time is essential in preventing the autolysis and necrosis of excised tissues and preserving their antigenicity. Five categories of fixatives are used in this stage: aldehydes, mercurials, alcohols, oxidizing agents, and picrates. The most common fixative used for imaging purposes is formaldehyde (also known as formalin), included in the aldehyde group. Fixation protocols are not standardized and vary according to the type of tissue and the histologic details needed to analyze it. The variability in this stage induces the possibility for several factors to affect this process, such as buffering (pH regulation), penetration (also depending on tissue thickness), volume (the usual ratio is 10:1), temperature, fixative concentration (10% solution is typical), and fixation time. These factors impact the quality of the scanned image, since stains used to highlight specific tissue properties may not react as expected.

After fixation, the tissue undergoes a processing stage necessary to create a paraffin embedding, which allows histotechnologists to cut the tissue into microscopic slides for further examination. The processing involves removing all water from the tissue using a series of alcohols and then clearing the tissue, which consists of removing the dehydrator with a miscible substance with the paraffin. Nowadays, tissue processors can automate this stage, by reducing inter-expert variability.

Dehydration and clearing will leave the tissue ready for the technician to create the embedded paraffin blocks. Depending on the tissue, these embeddings must be correctly aligned and oriented, determining which tissue section or cut is studied. Also, the embedding parameters (e.g., embedding temperature or peculiar chemicals involved) may defer from the norm for unique studies, so the research entity and the laboratory making the acquisition need to define them beforehand. Figure 1 shows a paraffin embedding cassette where the FFPE tissue samples can be stored even at room temperature for long periods.



Fig. 1 Paraffin cassettes

These embeddings undergo two more stages before being scanned: sectioning and staining. These procedures are discussed in the last section as they are no longer related to tissue preservation; instead, they are part of the tissue preparation stages before imaging.

2.2 Frozen Histological Tissue

Pathologists often use this tissue preservation method during surgical procedures where a rapid diagnosis of a pathological process is needed (extemporaneous preparation). In fact, frozen tissue produces the fastest stainable sections, although, compared to FFPE tissue, its morphological properties are not as good.

Frozen tissue (technically referred to as cryosection) is created by submerging the fresh tissue sample into cold liquid (e.g., pre-cooled isopentane in liquid nitrogen) or by applying a technique called *flash freezing*, which uses liquid nitrogen directly. As in FFPE, the tissue needs to be embedded in a medium to fix it to a chuck (i.e., specimen holder) in an optimal position for microscopic analysis. However, unlike FFPE tissue, no fixation or pre-embedding processes are needed for preservation.

For embedding, technicians use OCT (optimal cutting temperature compound), a viscous aqueous solution of polyvinyl alcohol and polyethylene glycol designed to freeze, providing the ideal support for cutting the cryosections in the cryostat (microtome under cold temperature). Different embedding approaches exist depending on the tissue orientation, precision and speed of the process, tissue wastage, and the presence of freeze artifacts in the resulting image. Stephen R. Peters describes these procedures and other important considerations needed to prepare tissue samples using the frozen technique [10].

Frozen tissue preservation relies on storing the embeddings at low temperatures. Therefore, the tissue will degrade if the cold chain breaks due to tissue sample mishandling. However, as it better preserves the tissue's molecular genetic material, it is frequently used in sequencing analysis and immunohistochemistry (IHC).

Other factors that affect the tissue quality and, therefore, the scanned images are the formation of ice crystals and the thickness of the sections. Ice crystals form when the tissue is not frozen rapidly enough, and it may negatively affect the tissue structure and, therefore, its morphological characteristics. On the other hand, frozen sections are often thicker than FFPE sections increasing the potential for lower resolution at higher magnifications and poorer images.

2.3 Tissue Preparation

We described the main pipeline to extract and preserve tissue samples for further analysis. Although the techniques described above can also be used for molecular and protein analysis (especially the frozen sections), we now focus only on the image pipeline by

describing the slide preparation for scanning and the potential artifacts observed in the acquired images.

Once the tissue embeddings are obtained, either by FFPE or frozen technique, they are prepared for viewing under a microscope or scanner. The tissue blocks are cut, mounted on glass slides, and stained with pigments (e.g., hematoxylin and eosin [H&E], saffron, or molecular biomarkers) to enhance the contrast and highlight specific cellular structures under the microscope.

Cutting the embeddings involves using a microtome to cut very thin tissue sections, later placed on the slide. The thickness of these sections is usually in the range of 4–20 microns. It will depend on the microscopy technique used for image acquisition and the experiment parameters. Special diamond knives are needed to get thinner sections, increasing the price of the microtome employed. If we use frozen embeddings, a cryostat keeps the environment's temperature low, avoiding tissue degradation.

Once on the slide, the tissue is heated to adhere to the glass and avoid wrinkles. If warming the tissue damages some of its properties (especially for immunohistochemistry), glue-coated slides can be used instead. For cryosections, pathologists often prefer to add a fixation stage to resemble the readings of an FFPE tissue section. This immediate fixation is achieved using several chemicals, including ethanol, methanol, formalin, acetone, or a combination. S. Peters describes the differences in the image quality based on these fixatives, as well as the proposed protocol for cutting and staining frozen sections [10]. For FFPE sections, Zhang and Xiong [11] describe neural histology's cutting, mounting, and staining methods. Protocols suggested by the authors are valuable guidelines for histotechnologists as tissue usually folds or tears, and bubbles form when cutting the embeddings. Minimizing these issues is essential to have good-quality images and accurate quantification of histological results.

Staining is the last process applied to the tissue before being imaged. Staining agents do not react with the embedding chemicals used to preserve the tissue sample; therefore, the tissue section needs to be cleaned and dried beforehand (e.g., eliminating all remains of paraffin wax used in the embedding). In [12], the authors present a review of the development of stains, techniques, and applications throughout time. One of the most common stains used in histopathology is hematoxylin and eosin (H&E). This agent highlights cell nuclei with a purple-blue color and the extracellular matrix and cytoplasm with the characteristic pink. Other structures in the tissue will show different hues, shades, and combinations of these colors. Figure 2 shows an H&E-stained human brainstem tissue and specific structures found on it.

Other staining agents can be used depending on the structure we would like to study or the clinical procedure. For instance, the toluidine blue stain is frequently used for intraoperative

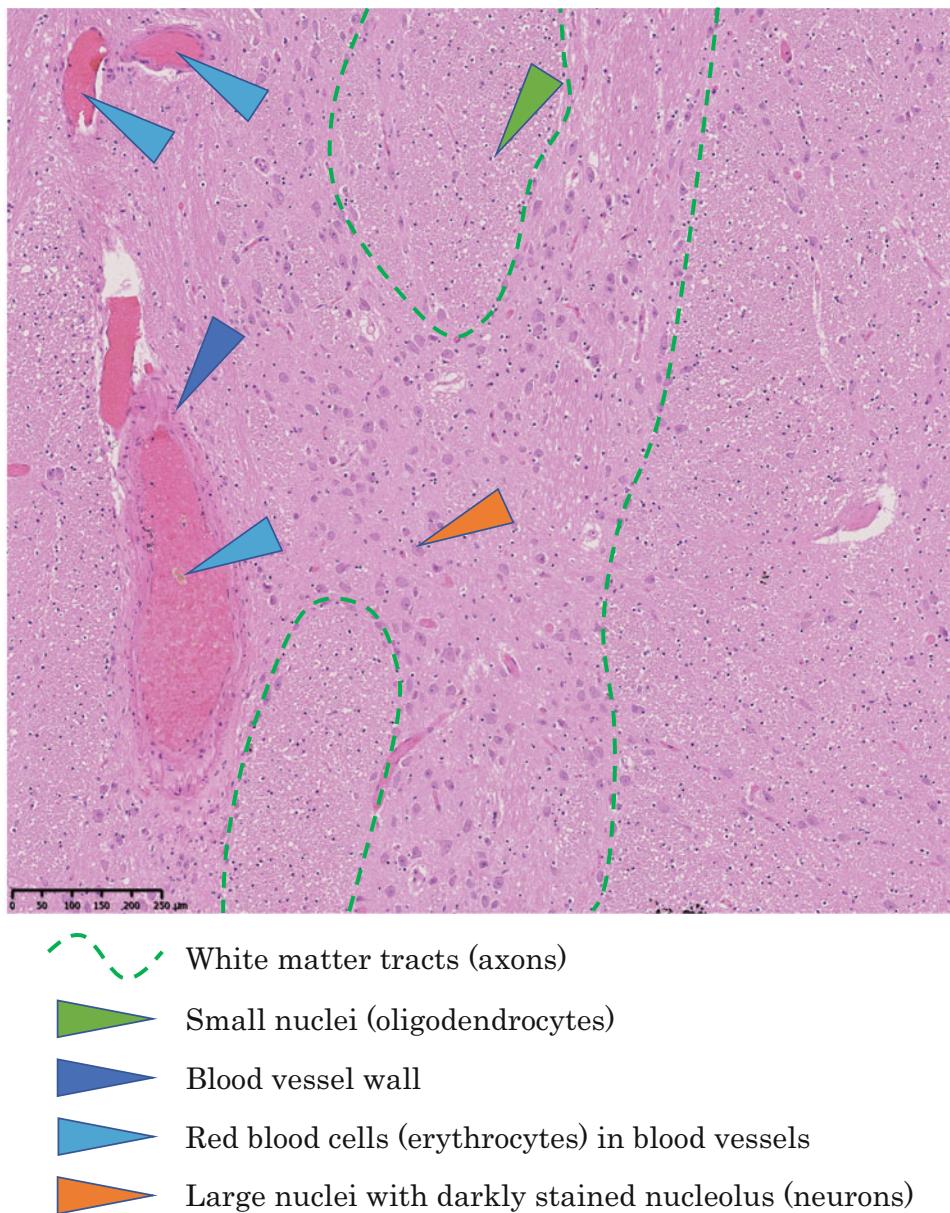


Fig. 2 H&E-stained WSI from human brainstem tissue preserved using FFPE. Relevant structures were annotated by expert pathologist. Abbreviations. H&E: hematoxylin and eosin. FFPE: formalin-fixed paraffin-embedded. WSI: whole slide image

consultation. Frozen sections are usually stained with this agent as it reacts almost instantly with the tissue. However, one disadvantage is that it only presents shades of blue and purple, so there is considerably less differential staining of the tissue structures [10].

For brain histopathology, other biomarkers are also available. For instance, the cresyl violet (or Nissl staining) is commonly used to identify the neuronal structure in the brain and spinal cord tissue

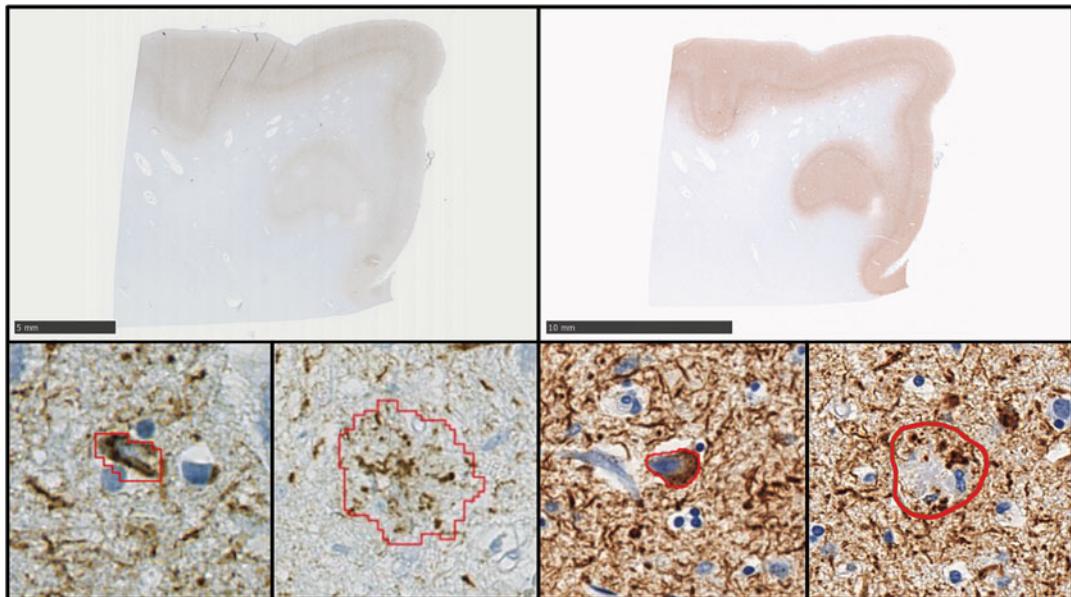


Fig. 3 [Top left] ALZ50 antibody used to discover compacted structures (tau pathologies). Below the WSI is an example of a neurofibrillary tangle (left) and a neuritic plaque (right) stained with ALZ50 antibody. [Top right] AT8 antibody, the most widely used in clinics, helps to discover all structures in a WSI. Below the WSI, there is an example of a neurofibrillary tangle (left) and a neuritic plaque stained with AT8 antibody (right). Abbreviation. WSI: whole slide image

[13]. Also, the Golgi method, which uses a silver staining technique, is used for observing neurons under the microscope [11]. Studies for Alzheimer's disease also frequently use ALZ50 and AT8 antibodies to reveal phosphorylated tau pathology using a standardized immunohistochemistry protocol [14–16]. Figure 3 shows the difference between ALZ50 and AT8 biomarkers and tau pathologies found in the tissue.

Having the slide stained is the last stage to prepare for studying microscopic structures of diseased or abnormal tissues. Considering the number of people involved in these processes (pathologists, pathology assistants, histotechnologists, tissue technicians, and trained repository managing personnel) and the precision of each stage, standardizing certain practices to create valuable slides for further analysis is needed.

Eiseman et al. [17] reported a list of best practices for biospecimen collection, processing, annotation, storage, and distribution. The proposal aims to set guidelines for managing large biospecimen banks containing the tissue sample embeddings excised from different organs with different pathologies and demographic distributions.

More specific standardized procedures for tissue sampling and processing have also been reported. For instance, in 2012, the Society of Toxicologic Pathology charged a Nervous System Sampling Working Group with devising recommended practices to

routinely screen the central nervous system (CNS) and peripheral nervous system (PNS) during nonclinical general toxicity studies. The authors proposed a series of approaches and recommendations for tissue fixation, collection, trimming, processing, histopathology examination, and reporting [18]. Zhang J. et al. also address the process of tissue preparation, sectioning, and staining but focus only on brain tissue [11]. Although these recommendations aim to standardize specific techniques among different laboratories, they are usually imprecise and approximate, leaving the final decision to the specialists based on the tissue handled.

Due to this lack of automation during surgical removal, fixation, tissue processing, embedding, microtomy, staining, and mounting procedures, several artifacts can impact the quality of the image and the results of the analysis. A review of these artifacts is presented in [19]. The authors review the causes of the most frequent artifacts, how to identify them, and propose some ideas to prevent them from interfering with the diagnosis of lesions. For better understanding and following the tissue preparation and image acquisition procedure, the authors proposed a classification of eight classes: prefixation artifacts, fixation artifacts, artifacts related to bone tissue, tissue-processing artifacts, artifacts related to microtomy, artifacts related to floatation and mounting, staining artifacts, and mounting artifacts. Figure 4 shows some of them.

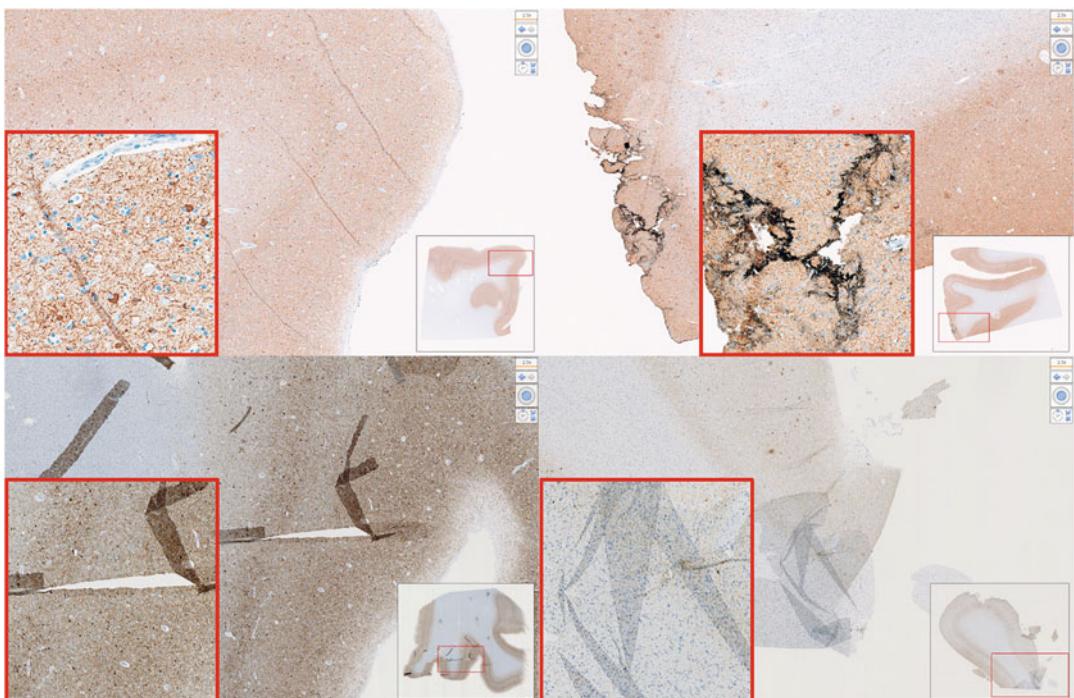


Fig. 4 [Top left] Folding artifact (floatation and mounting-related artifact), [Top right] Marking fixation process (fixation artifact), [Bottom left] Breaking artifact (microtome-related artifact), [Bottom right] Overlaying tissue (mounting artifact)

3 Histopathological Image Analysis

This section aims to better understand the role that digital pathology plays in the analysis of complex and large amounts of information obtained from tissue specimens. As an additional option to incorporate more images with higher throughput, whole slide image scanners are briefly discussed. Therefore, we must discuss the DICOM standard used in medicine to digitally represent the images and, in this case, the tissue samples. We then focus on computational pathology, which is the analysis of the reconstructed whole slide images using different pattern recognition techniques such as machine learning (including deep learning) algorithms. This section contains some extractions from Jiménez's thesis work [20].

3.1 Digital Pathology

Digital systems were introduced to the histopathological examination in order to deal with complex and vast amounts of information obtained from tissue specimens. Digital images were originally generated by mounting a camera on the microscope. The static pictures captured only reflected a small region of the glass slide, and the reconstruction of the whole glass slide was not frequently attempted due to its complexity and the fact that it is time-consuming. However, precision in the development of mechanical systems has made possible the construction of whole slide digital scanners. Garcia et al. [21] reviewed a series of mechanical and software systems used in the construction of such devices. The stored high-resolution images allow pathologists to view, manage, and analyze the digitized tissue on a computer monitor, similar to under an optical microscope but with additional digital tools to improve the diagnosis process.

WSI technology, also referred to as *virtual microscopy*, has proven to be helpful in a wide variety of applications in pathology (e.g., image archiving, telepathology, image analysis). In essence, a WSI scanner operation principle consists of moving the glass slide a small distance every time a picture is taken to capture the entire tissue sample. Every WSI scanner has six components: (a) a microscope with lens objectives, (b) a light source (bright field and/or fluorescent), (c) robotics to load and move glass slides around, (d) one or more digital cameras for capture, (e) a computer, and (f) software to manipulate, manage, and view digital slides [22]. The hardware and software used for these six components will determine the key features to analyze when choosing a scanner. Some research articles have compared the hardware and software capabilities of different scanners in the market. For instance, in [22], Farahani et al. compared 11 WSI scanners from different manufacturers regarding imaging modality, slide capacity, scan speed, image magnification, image resolution, digital slide format,

multilayer support, and special features their hardware and software may offer. This study showed that robotics and hardware used in a WSI scanner are currently state of the art and almost standard in every device. Software, on the other hand, has some ground for further development. A similar study by Garcia et al. [21] reviewed 31 digital slide systems comparing the same characteristics in Farahani's work. In addition, the authors classified the devices into digital microscopes (WSI) for virtual slide creation and diagnosis-aided systems for image analysis and telepathology. Automated microscopes were also included in the second group as they are the baseline for clinical applications.

3.2 Whole Slide Image Structure

The Digital Imaging and Communications in Medicine (DICOM) standard was adopted to store WSI digital slides into commercially available PACS (picture archiving and communication system) and facilitate the transition to digital pathology in clinics and laboratories. Due to the WSI dimension and size, a new pyramidal approach for data organization and access was proposed by the DICOM Standards Committee in [23].

A typical digitalization of a $20\text{ mm} \times 15\text{ mm}$ sample using a resolution of $0.25\text{ }\mu\text{m/pixel}$, also referred to as $40\times$ magnification, will generate an image of approximately $80,000 \times 60,000$ pixels. Considering a 24-bit color resolution, the digitized image size is about 15 GB. Data size might even go one order of magnitude higher if the scanner is configured to a higher resolution (e.g., $80\times$, $100\times$), Z planes are used, or additional spectral bands are also digitized. In any case, conventional storage and access to these images will demand excessive computational resources to be implemented into commercial systems. Figure 5 describes the traditional approach (i.e., *single frame* organization), which stores the data in rows that extend across the entire image. This row-major approach has the disadvantage of loading unnecessary pixels into memory, especially if we want to visualize a small region of interest.

Other types of organizations have also been studied. Figure 6 describes the storage of pixels in *tiles*, which decreases the computational time for visualization and manipulation of WSI by loading only the subset of pixels needed into memory. Although this approach allows faster access and rapid visualization of the WSI, it fails when dealing with different magnifications of the images, as is the case in WSI scanners. Figure 7 depicts the issues with rapid zooming of WSI. Besides loading a larger subset of pixels into memory, algorithms to perform the down-sampling of the image are time-consuming. At the limit, to render a low-resolution thumbnail of the entire image, all the data scanned must be accessed and processed [23]. Stacking precomputed low-resolution versions of the original image was proposed in order to overcome the zooming problem. Figure 8 describes the *pyramidal* structure used to store different down-sampled versions

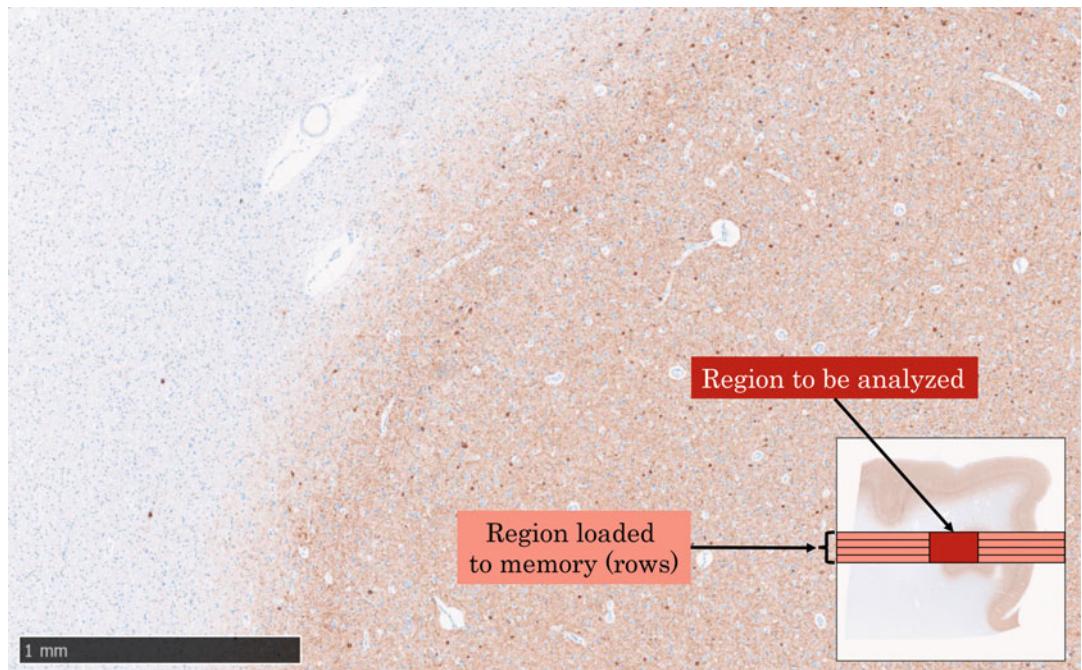


Fig. 5 Single frame organization of whole slide images

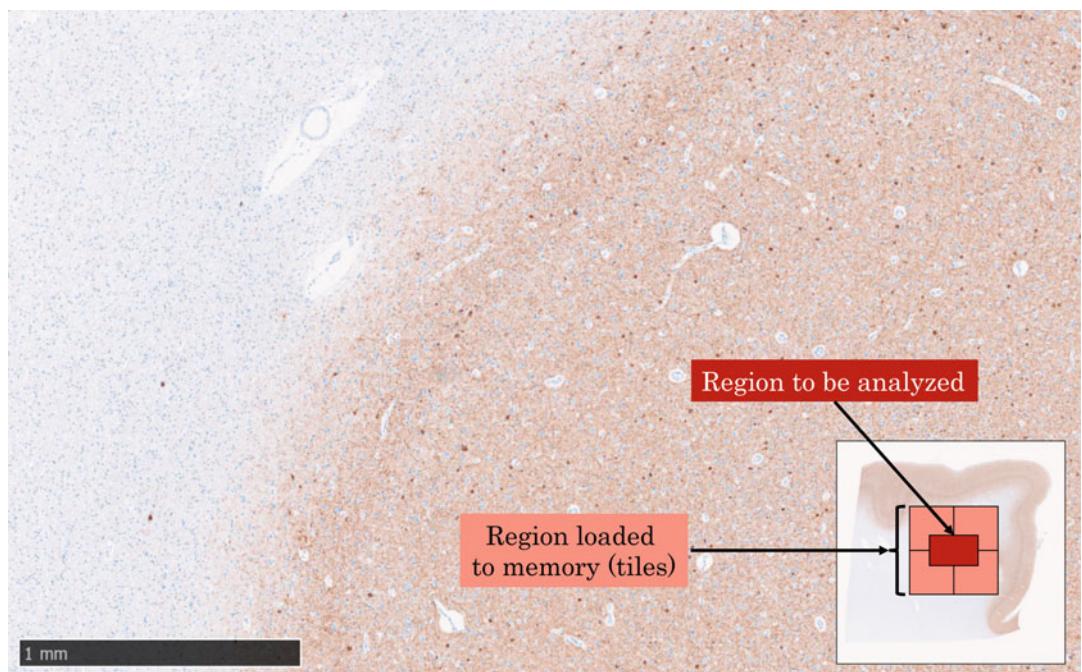


Fig. 6 Tiled image organization of whole slide images. Tiles' size can range from 240×240 pixels up to 4096×4096 pixels

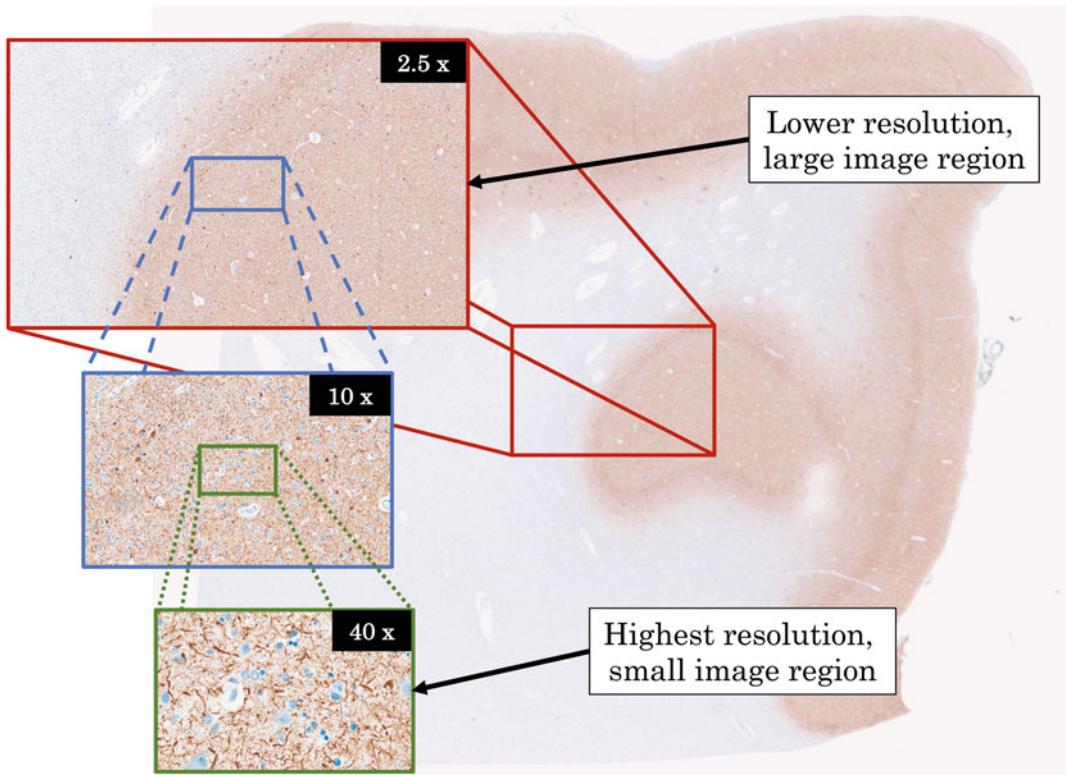


Fig. 7 Rapid zooming issue when accessing lower-resolution images: large amount of data need to be loaded into memory. In this example, the image size at the highest resolution (221 nm/pixels) is 82,432 × 80,640 pixels

of the original image. The bottom of the pyramid corresponds to the highest resolution and goes up to the thumbnail (lowest resolution) image. For further efficiency, tiling and pyramidal methods are combined to facilitate rapid retrieval of arbitrary subregions of the image as well as access to different resolutions. As depicted in Fig. 8, each image in the pyramid is stored as a series of tiles. In addition, the baseline image tiles can contain different colors or z-planes if multispectral images are acquired or if tracking variations in the specimen thickness are needed. This combined approach can be easily integrated into a web architecture such as the one presented by Lajara et al. [24] as tiles of the current user's viewport can be cached without high memory impact.

As mentioned in previous paragraphs, WSI can occupy several terabytes of memory due to the data structure. Depending on the application, lossless or lossy compression algorithms can be applied. Lossless compression typically yields a 3X–5X reduction in size; meanwhile, lossy compression techniques such as JPEG and JPEG2000 can achieve from 15X–20X up to 30X–50X reduction, respectively [23]. Due to no standardization of WSI file formats,

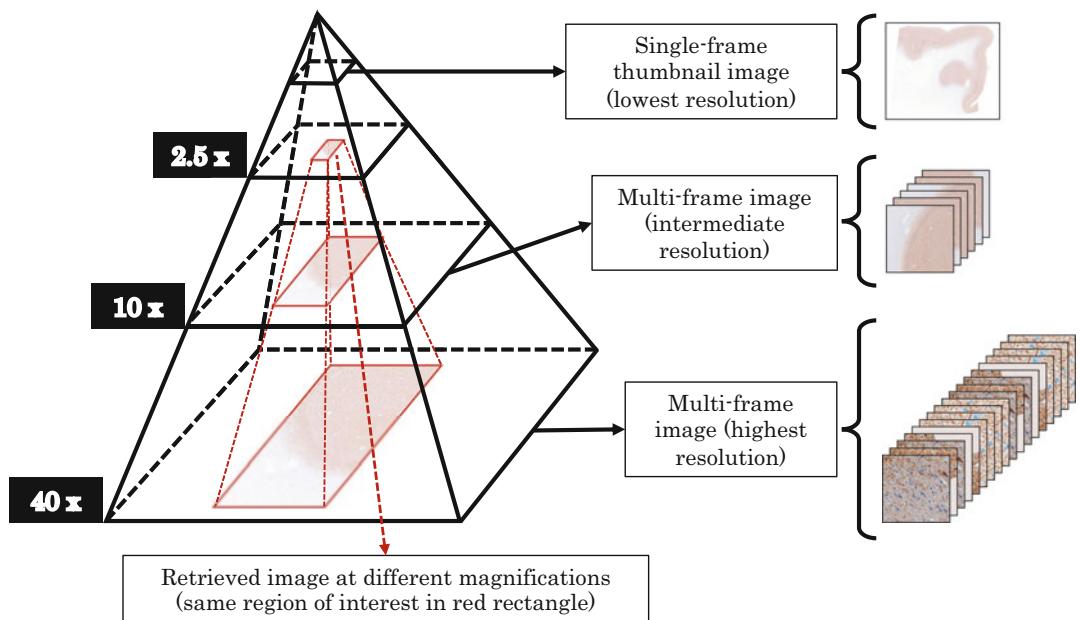


Fig. 8 Pyramidal organization of whole slide images. In this example, the image size at the highest resolution (221 nm/pixels) is 82,432 × 80,640 pixels. The compressed (JPEG) file size is 2.22 GB, whereas the uncompressed version is 18.57 GB

scan manufacturers may also develop their proprietary compression algorithms based on JPEG and JPEG2000 standards. Commercial WSI formats have a mean default compression value ranging from 13X to 27X. Although the size of WSI files is considerably reduced, efficient data storage was not the main issue when designing WSI formats for more than 10 years. In [25], Helin et al. addressed this issue and proposed an optimization to the JPEG2000 format, which yields up to 176X compression. Although no computational time has been reported in the aforementioned study, this breakthrough allows for efficient transmission of data through systems relying on Internet communication protocols.

3.3 Computational Pathology

Computational pathology is a term that refers to the integration of WSI technology and image analysis tools in order to perform tasks that were too cumbersome or even impossible to undertake manually. Image processing algorithms have evolved, yielding enough precision to be considered in clinical applications, such is the case for surgical pathology using frozen samples reported by Bauer et al. in [26]. Other examples mentioned in [22] include morphological analysis to quantitatively measure histological structures [27], automated selection of regions of interest such as areas of most active proliferative rate [28], and automated grading of tumors [29]. Moreover, educational activities have also benefited from the development of computational pathology. Virtual tutoring,

online medical examinations, performance improvement programs, and even interactive *illustrations* in articles and books are being implemented, thanks to this technology [22].

In order to validate a WSI scanner for clinical use (diagnosis purposes), several tests are conducted following the guidelines developed by the College of American Pathologists (CAP) [30]. On average, reported discrepancies between digital slides and glass slides are in the range of 1–5%. However, even glass-to-glass slide comparative studies can yield discrepancies due to observer variability, and increasing case difficulty.

Although several studies in the medical community have reported using WSI scanners to perform the analysis of tissue samples, pathologists remain reluctant to adopt this technology in their daily practice. Lack of training, limiting technology, shortcomings in scanning all slides, cost of equipment, and regulatory barriers have been reported as the principal issues [22]. In fact, it was until early 2017 that the first WSI scanner was approved by the FDA and released to the market [31]. Nevertheless, WSI technology represents a milestone in modern pathology, having the potential to enhance the practice of pathology by introducing new tools which help pathologists provide a more accurate diagnosis based on quantitative information. Besides, this technology is also a bridge for bringing omics closer to routine histopathology toward future breakthroughs as spatial transcriptomics.

4 Methods in Brain Computational Pathology

This section is dedicated to different machine learning and deep learning methodologies to analyze brain tissue samples. We describe the technology by focusing on how this is applied (i.e., at the WSI or the patch level), the medical task associated with it, the dataset used, the core structure/architecture of the algorithms, and the significant results.

We begin by describing the general challenges in WSI analysis. Then we move on to deep learning methods concerning only WSI analysis, and we finalize with machine learning and deep learning applications for brain disorders focusing on the disease rather than the processing of the WSI. In addition, as in the primary biomedical areas, data annotation is a vital issue in computational pathology, generating accurate and robust results. Therefore, some new techniques used to create reliable annotations—based on a seed-annotated dataset—will be presented and discussed.

4.1 Challenges in WSI Analysis Using ML

Successful application of machine learning algorithms to WSIs can improve—or even surpass—the accuracy, reproducibility, and objectivity of current clinical approaches and propel the creation of new clinical tools providing new insights on various pathologies

[32]. Due to the characteristics of a whole slide image and the acquisition process described in the sections above, researchers usually face two nontrivial challenges related to the visual understanding of the WSIs and the inability of hardware and software to facilitate learning from such high-dimensional images.

Regarding the first challenge, the issue relies on the lack of generalization of ML techniques due to image artifacts and color variability in staining. Imaging artifacts directly result from the tissue section processing errors and the hardware (scanners) used to digitize the slide. The uneven illumination, focusing, and image tiling are a few imaging artifacts present in the WSI, being the first the most relevant and studied as it is challenging for an algorithm to extract useful features from some regions of the scanned tissue. It gets even worse when staining artifacts such as stain variability are also present.

To address this problem, we find several algorithms for color normalization in the literature. Macenko [33], Vahadane [34], and Reinhard [35] algorithms are classical algorithms for color normalization implementing image processing techniques such as histogram normalization, color space transformations, color deconvolution (color unmixing), reference color density maps, or histogram matching. Extensions from these methods are also reported. For instance, Magee et al. [36] proposed two approaches to extend the Reinhard method: a multimodal linear normalization in the *Lab* color space and normalization in a representation space using stain-specific color deconvolution.

The use of machine learning techniques, specifically deep convolutional neural networks, has also been studied for color normalization. In [37], the authors proposed the StainNet for stain normalization. The framework consists of a GAN² (*teacher* network) trained to learn the mapping relationship between a source and target image, and an FCNN³ (*student* network) able to transfer the mapping relationship of the GAN based on image content into a mapping relationship based on pixel values. A similar approach using cycle-consistent GANs was also proposed for the normalization of H&E-stained WSIs [38]. In the last case, synthetically generated images capture the representative variability in the color space of the WSI, enabling the architecture to transfer any color information from a new source image into a target color space.

On the other hand, the second challenge related to the high dimensionality of WSIs is addressed in two ways: processing using patch-level or slide-level annotations. Dimitriou N. et al. reported an overview of the literature for both approaches in [32]. For patch-based annotations, the authors reported patch sizes ranging

² GAN: generative adversarial networks.

³ FCNN: fully convolutional neural network.

from 32×32 pixels up to $10,000 \times 10,000$ pixels and a frequent value of 256×256 pixels. Patches are generated and processed by sequentially dividing the WSI into tiles, which demand higher computational resources, by random sampling, leading to class imbalance issues, or by following a guided sampling based on pixel annotations. Patch-level annotations usually contain pixel-level labels. Frequently approaches using these annotations focus on the segmentation of morphological structures in patches rather than the classification of the entire WSI. In [39], the authors studied the potential of semantic architectures such as the U-Net and compared it to classical CNN approaches for pixel-wise classification. Another approach known as HistoSegNet [40] implements a combination of visual attention maps (or activation maps) using the Grad-CAM algorithm and CNN for semantic segmentation of WSIs. In addition, several methods are summarized in [41, 42] using graph deep neural networks to detect and segment morphological structures in WSIs.

Pixel labeling at high resolution is a time-demanding task and is prone to inter- and intra-expert variabilities impacting the learning process of machine learning algorithms. Therefore, despite the lower granularity of labeling, several studies have shown promising results when working with slide-based annotations.

With no available information about the pixel label, most algorithms usually aim to identify patches (or regions of interest in the WSI) that can collectively or independently predict the classification of the WSI. These techniques often rely on multiple instance learning, unsupervised learning, reinforcement learning, transfer learning, or a combination thereof [32]. Tellez et al. [43] proposed a two-step method for gigapixel histopathology analysis based on an unsupervised neural network compression algorithm to extract latent representations of patches and a CNN to predict image-level labels from those compressed images. In [44], the authors proposed a four-stage methodology for survival prediction based on randomly sampled patches from different patients' slides. They used PCA to reduce the features' space dimension prior to the K-means clustering process to group patches according to their phenotype. Then, a deep convolutional network (DeepConvSurv) is used to determine which patches are relevant for the aggregation and final survival score. Qaiser et al. [45] proposed a model mimicking the histopathologist practice using recurrent neural networks (RNN) and CNN. In their proposal, they treat images as the *environment* and the RNN+CNN as the *agent* acting as a decision-maker (same as the histopathologists). The agent then looks at high-level tissue components (low magnification) and evaluates different regions of interest at low-level magnification, storing relevant morphological features into *memory*. Similarly, Momeni et al. [46] suggested using deep recurrent attention models (DRAMs) and CNN to create an attention-based architecture to

process large input patches and locate discriminatory regions more efficiently. This last approach needs, however, further validation as results are not conclusive and have not been accepted by the scientific community yet.

Relevant features for disease analysis, diagnosis, or patient stratification can be extracted from individual patches by looking into cell characteristics or morphology; however, higher structural information, such as the shape or extent of a tumor, can only be captured in more extensive regions. Some approaches to processing multiple magnification levels of a WSI are reported in [47–51]. They involve leveraging the pyramidal structure of WSI to access features from different resolutions and model spatial correlations between patches.

All the studies cited so far have no specific domain of application. Most of them were trained and tested using synthetic or public datasets containing tissue pathologies from different body areas. Therefore, most of the approaches can extend to different pathologies and diseases. In the following subsections, however, we will focus only on specific brain disorder methodologies.

4.2 DL Algorithms for Brain WSI Analysis

In recent times, deep-learning-based methods have shown promising results in digital pathology [52]. Unfortunately, only a few public datasets contain WSI of brain tissue, and most of them only contain brain tumors. In addition, most of them are annotated at the slide level, making the semantic segmentation of structures more challenging. Independently of the task (i.e., detection/classification or segmentation) and the application in brain disorders, we will explore the main ideas behind the methodologies proposed in the literature.

For the analysis of benign or cancerous pathologies in brain tissue, tumor cell nuclei are of significant interest. The usual framework for analyzing such pathologies was reported in [53] and used the WSI of diffuse glioma. The method first segments the regions of interest by applying classical image processing techniques such as mathematical morphology and thresholding. Then, several hand-crafted features such as nuclear morphometry, region texture, intensity, and gradient statistics were computed and inputted to a nuclei classifier. Although such an approach—using quadratic discriminant analysis and maximum a posteriori (MAP) as a classification mechanism—reported an overall accuracy of 87.43%, it falls short compared to CNN, which relies on automated feature extractions using convolutions rather than on handcrafted features. Xing et al. [54] proposed an automatic learning-based framework for robust nucleus segmentation. The method begins by dividing the image into small regions using a sliding window technique. These patches are then fed to a CNN to output probability maps and generate initial contours for the nuclei using a region merging algorithm. The correct nucleus segmentation is obtained by

alternating dictionary-based shape deformation and inference. This method outperformed classical image processing algorithms with promising results (mean Dice similarity coefficient of 0.85 and detection F_1 score of 0.77 computed using gold-standard regions within 15 pixels for every nucleus center) using CNN-based features over classical ones.

Following a similar approach, Xu et al. [55] reported the use of deep convolutional activation features for brain tumor classification and segmentation. The authors used a pre-trained AlexNet CNN [56] on the ImageNet dataset to extract patch features from the last hidden layer of the architecture. Features are then ranked based on the difference between the two classes of interest, and the top 100 are finally input to an SVM for classification. For the segmentation of necrotic tissue, an additional step involving probability mappings from SVM confidence scores and morphological smoothing is applied. Other approaches leveraging the use of CNN-based features for glioma are presented in [47, 57]. The experiments reported achieved a maximum accuracy of 97.5% for classification and 84% for segmentation. Although these results seemed promising, additional tests with different patch sizes in [47] suggested that the method's performance is data-dependent as numbers increase when larger patches, meaning more context information, are used.

With the improvement of CNN architectures for natural images, more studies are also leveraging transfer learning to propose end-to-end methodologies for analyzing brain tumors. Ker et al. [58] used a pre-trained Google Inception V3 network to classify brain histology specimens into normal, low-grade glioma (LGG), or high-grade glioma (HGG). Meanwhile, Truong et al. [59] reported several optimization schemes for a pre-trained ResNet-18 for brain tumor grading. The authors also proposed an explainability tool base on tile-probability maps to aid pathologists in analyzing tumor heterogeneity. A summary of DL approaches used in brain WSI processing, alongside other brain imaging modalities such as MRI or CT, is reported by Zadeh et al. in [60].

Let us now focus on studies dealing with tau pathology, which is a hallmark of Alzheimer's disease. In [61], three different DL models were used to segment tau aggregates (tangles) and nuclei in postmortem brain WSIs of patients with Alzheimer's disease. The three models included an FCNN, U-Net [62], and SegNet [63], with SegNet achieving the highest accuracy in terms of the intersection-over-union index. In [64], an FCNN was used on a dataset of 22 WSIs for semantic segmentation of tangle objects from postmortem brain WSIs. Their model is able to segment tangles of varying morphologies with high accuracy under diverse staining intensities. An FCNN model is also used in [65] to classify morphologies of tau protein aggregates in the gray and white

matter regions from 37 WSIs representing multiple degenerative diseases. In [14], tau aggregate analysis is processed on a dataset of six postmortem brain WSIs with a combined classification-segmentation framework which achieved an F_1 score of 81.3% and 75.8% on detection and segmentation tasks, respectively. In [16], neuritic plaques have been processed from eight human brain WSIs from the frontal lobe, stained with AT8 antibody (majorly used in clinics, helping to highlight most of the relevant structures). The impact of the staining (ALZ50 [14] vs. AT8 [16]), the normalization method, the slide scanner, the context, and the DL traceability/explainability have been studied, and a comparison with commercial software has been made. A baseline of 0.72 for the Dice score has been reported for plaque segmentation, reaching 0.75 using an attention U-Net.

Several domains in DL-based histopathological analysis of AD tauopathy remain unexplored. Firstly, even if, as discussed, a first work concerning neuritic plaques has been recently published by our team in [16], most of the existing works have used DL for segmentation of tangles rather than plaques, as the latter are harder to identify against the background gray matter due to their diffuse/sparse appearance. Secondly, annotations of whole slide images are frequently affected by errors by human annotators. In such cases, a DL preliminary model may be trained using weakly annotated data and used to assist the expert in refining annotations. Thirdly, contemporary tau segmentation studies do not consider context information. This is important in segmenting plaques from brain WSIs as these occur as sparse objects against an extended background of gray matter. Finally, DL models with explainability features have not yet been applied in tau segmentation from WSIs. This is a critical requirement for DL models used in clinical applications [66] [67]. The DL models should not only be able to precisely identify regions of interest, but clinicians and general users need to know the discriminative image features the model identifies as influencing their decisions.

4.3 Applications of Brain Computational Pathology

Digital systems were introduced to the histopathological examination to deal with complex and vast amounts of information obtained from tissue specimens. Whole slide imaging technology has proven to be helpful in a wide variety of applications in pathology (e.g., image archiving, telepathology, image analysis), especially when combining this imaging technique with powerful machine learning algorithms (i.e., computational pathology).

In this section, we will describe some applications of computational pathology for the analysis of brain tissue. Most methods focus on tumor analysis and cancer; however, we also find interesting results in clinical applications, drug trials [68], and neurodegenerative diseases. The authors cited in this section aim to understand brain disorders and use deep learning algorithms to extract relevant information from WSI.

In brain tumor research, an early survival study for brain glioma is presented in [44]. The approach has been previously described above. In brief, it is a four-stage methodology based on randomly sampled patches from different patients' slides. They perform dimensionality reduction using PCA and then K-means clustering to group patches according to their phenotype. Then, the patches are sent to a deep convolutional network (DeepConvSurv) to determine which were relevant for the aggregation and final survival score. The deep survival model is trained on a small dataset leveraging the architecture the authors proposed. Also, the method is annotation-free, and it can learn information about one patient, regardless of the number or size of the WSIs. However, it has a high computational memory footprint as it needs hundreds of patches from a single patient's WSI. In addition, the authors do not address the evaluation of the progression of the tumor, and a deeper analysis of the clusters could provide information about the phenotypes and their relation to brain glioma.

Whole slide images have been used as a primary source of information for cancer diagnosis and prognosis, as they reveal the effects of cancer onset and its progression at the subcellular level. However, being an invasive image modality (i.e., tissue gathered during a biopsy), it is less frequently used in research and clinical settings. As an alternative, noninvasive and nonionizing imaging modalities, such as MRI, are quite popular for oncology imaging studies, especially in brain tumors.

Although radiology and pathology capture morphologic data at different biological scales, a combination of image modalities can improve image-based analysis. In [69], the authors presented three classification methods to categorize adult diffuse glioma cases into oligodendrogloma and astrocytoma classes using radiographic and histologic image data. Thirty-two cases were gathered from the TCGA project⁴ containing a set of MRI data (T1, T1C, FLAIR, and T2 images) and its corresponding WSI, taken from the same patient at the same time point. The methods described were proposed in the context of the Computational Precision Medicine (CPM) satellite event at MICCAI 2018, one of the first combining radiology and histology imaging analyses. The first one develops two independent pipelines giving two probability scores for the prediction of each case. The MRI pipeline preprocesses all images to remove the skull, co-register, and resample the data to leverage a fully convolutional neural network (CNN) trained on another MRI dataset (i.e., BraTS-2018) to segment tumoral regions. Several radiomic features are computed from such regions, and after reducing its dimensionality with PCA, a logistic regression classifier

⁴ <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/typesan>.

outputs the first probability score. WSIs also need a preprocessing stage as tissue samples may contain large areas of glass background. After a color space transformation to HSV (hue saturation value), lower and upper thresholds are applied to get a binary mask with the region of interest, which is then refined using mathematical morphology. Color-normalized patches of 224×224 pixels are extracted from the region of interest (ROI) and filtered to exclude outliers. The remaining patches are used to refine a CNN (i.e., DenseNet-161) pre-trained on the ImageNet dataset. In the prediction phase, the probability score of the WSI is computed using a voting system of the classes predicted for individual patches. The scores from both pipelines are finally processed in a confidence-based voting system to determine the final class of each case. This proposal achieved an accuracy score of 0.9 for classification.

The second and third approaches also processed data in two different pipelines. There are slight variations in the WSI preprocessing step in the second method, including Otsu thresholding for glass background removal and histogram equalization for color normalization of patches of 448×448 pixels. Furthermore, the authors used a 3D CNN to generate the output predictions for the MRI data and a DenseNet pre-trained architecture for WSI patch classification. The last feature layer from each classification model is finally used as input to an SVM model for a unified prediction. In addition, regularization using dropout is performed in the test phase to avoid overfitting the models. The accuracy obtained with this methodology was 0.8.

The third approach uses larger patches from WSI and an active learning algorithm proposed in [70] to extract regions of interest instead of randomly sampling the tissue samples. Features from the WSI patches are extracted using a VGG16 CNN architecture. The probability score is combined with the output probability of a U-Net + 2D DenseNet architecture used to process the MRI data. The method achieved an accuracy of 0.75 for unified classification. Although results are promising and provide a valid approach to combining imaging modalities, data quality and quantity are still challenging. The use of pre-trained CNN architectures for transfer learning using a completely different type of imaging modality might impact the performance of the whole pipeline. As seen in previous sections, WSI presents specific characteristics depending on the preparation and acquisition procedures not represented in the ImageNet dataset.

An extension to the previous study is presented in [71]. The authors proposed a two-stage model to classify gliomas into three subtypes. WSIs were divided into tiles and filtered to exclude patches containing glass backgrounds. An ensemble learning framework based on three CNN architectures (EfficientNet-B2, EfficientNet-B3, and SEResNeXt101) is used to extract features which are then combined with meta-data (i.e., age of the patient) to

predict the class of glioma. MRI data is preprocessed in the same way as described before and input to a 3D CNN network with a 3D ResNet architecture as a backbone.

The release of new challenges and datasets, such as the Computational Precision Medicine: Radiology-Pathology Challenge on brain tumor classification (CPM-RadPath), has also allowed studies using weakly supervised deep learning methods for glioma subtype classification. For instance, in [72], the authors combine 2D and 3D CNN to process 388 WSI, and its corresponding multiparametric MRI collected from the same patients. Based on a confidence index, the authors were able to fuse WSI- and-MRI-based predictions improving the final classification of the glioma subtype.

Moving on from brain tumors, examining brain WSI also provides essential insights into spatial characteristics helpful in understanding brain disorders.

In this area, analyzing small structures present in postmortem brain tissue is crucial to understanding the disease deeply. For instance, in Alzheimer's disease, tau proteins are essential markers presenting the best histopathological correlation with clinical symptoms [73]. Moreover, these proteins can aggregate in three different structures within the brain (i.e., neurites, tangles, and neuritic plaques) and constitute one significant biomarker to study the progression of the disease and stratify patients accordingly.

In [14], the authors addressed the detection task of the Alzheimer's patient stratification pipeline. The authors proposed a U-Net-based methodology for tauopathies segmentation and a CNN-based architecture for tau aggregates' classification. In addition, the pipelines were completed with a nonlinear color normalization preprocessing and a morphological analysis of segmented objects. These morphological features can aid in the clustering of patients having different disease manifestations. One limitation, however, is the accuracy obtained in the segmentation/detection process.

Understanding the accumulation of abnormal tau protein in neurons and glia allows differentiating tauopathies such as Alzheimer's disease, progressive supranuclear palsy (PSP), cortico-basal degeneration (CBD), and Pick's disease (PiD). In [74], the authors proposed a diagnostic tool consisting of two stages: (1) an object detection pipeline based on the CNN YOLOv3 and (2) a random forest classifier. The goal is to detect different tau lesion types and then analyze their characteristics to determine to which specific pathology they belong. With an accuracy of 0.97 over 2522 WSI, the study suggests that machine learning methods can be applied to help differentiate uncommon neurodegenerative tauopathies.

Tauopathies are analyzed using postmortem brain tissue samples. For in vivo studies, there exist tau PET tracers that, unfortunately, have not been validated and approved for clinical use as

correlations with histological samples are needed. In [75], the authors proposed an end-to-end solution for performing large-scale, voxel-to-voxel correlations between PET and high-resolution histological signals using open-source resources and MRI as the common registration space. A U-Net-based architecture segments tau proteins in WSI to generate 3D tau inclusion density maps later registered to MRI to validate the PET tracers. Although segmentation performance was around 0.91 accurate in 500 WSI, the most significant limitation is the tissue sample preparation, meaning extracting and cutting brain samples to reconstruct 3D histological volumes. Additional studies combining postmortem MRI and WSI for neurodegenerative diseases were reported by Jonkman et al. in [76].

5 Perspectives

This last section of the chapter deals with new techniques for the explainability of artificial intelligence algorithms. It also describes new ideas related to responsible artificial intelligence in the context of medical applications, computational histopathology, and brain disorders. Besides, it introduces new image acquisition technology mixing bright light and chemistry to improve intraoperative applications. Finally, we will highlight computational pathology's strategic role in spatial transcriptomics and refined personalized medicine.

In [15, 16], we address the issue of accurate segmentation by proposing a two-loop scheme as shown in Fig. 9. In our method, a U-Net-based neural network is trained on several WSIs manually annotated by expert pathologists. The structures we focus on are neuritic plaques and tangles following the study in [14]. The network's predictions (in new WSIs) are then reviewed by an expert who can refine the predictions by modifying the segmentation outline or validating new structures found in the WSI. Additionally, an attention-based architecture is used to create a visual explanation and refine the hyperparameters of the initial architecture in charge of the prediction proposal.

We tested the attention-based architecture with a dataset of eight WSIs divided into patches following an ROI-guided sampling. Results show qualitatively in Fig. 10 that through this visual explanation, the expert in the loop could define the border of the neuritic plaque (object of interest) more accurately so the network can update its weights accordingly. Additionally, quantitative results (Dice score of approximately 0.7) show great promise for this attention U-Net architecture.

Our next step is to use a single architecture for explainability and segmentation/classification. We believe our method will improve the accuracy of the neuritic plaques and tangles outline

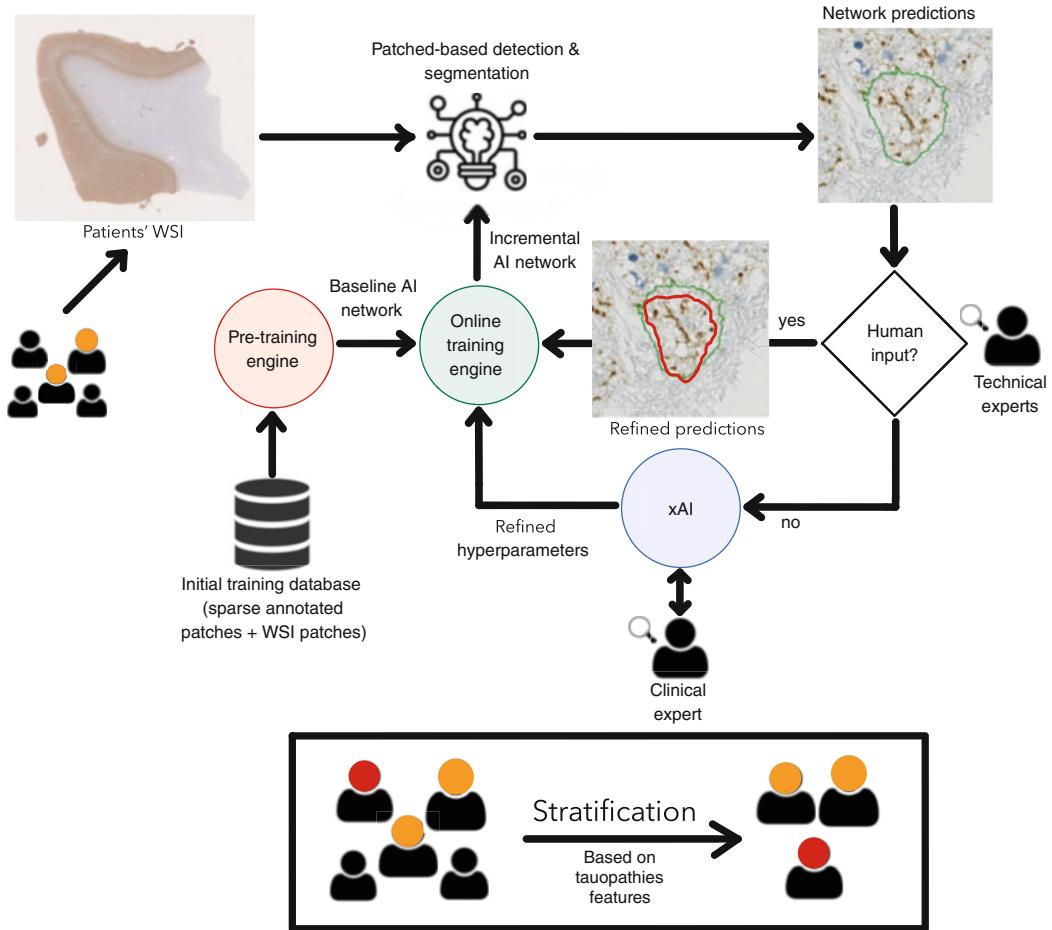


Fig. 9 Expert-in-the-loop architecture proposal to improve tauopathies' segmentation and to stratify AD patients

and create better morphological features for patient stratification and understanding of Alzheimer's disease [15, 16].

Despite their high computational efficiency, artificial intelligence—in particular deep learning—models face important usability and translational limitations in clinical use, as in biomedical research. The main reason for these limitations is generally low acceptability by biomedical experts, essentially due to the lack of feedback, traceability, and interpretability. Indeed, domain experts usually feel frustrated by a general lack of insights, while the implementation of the tool itself requires them to make a considerable effort to formalize, verify, and provide a tremendous amount of domain expertise. Some authors speak of a “black-box” phenomenon, which is undesirable for a traceable, interpretable, explicable, and ultimately responsible use of these tools.

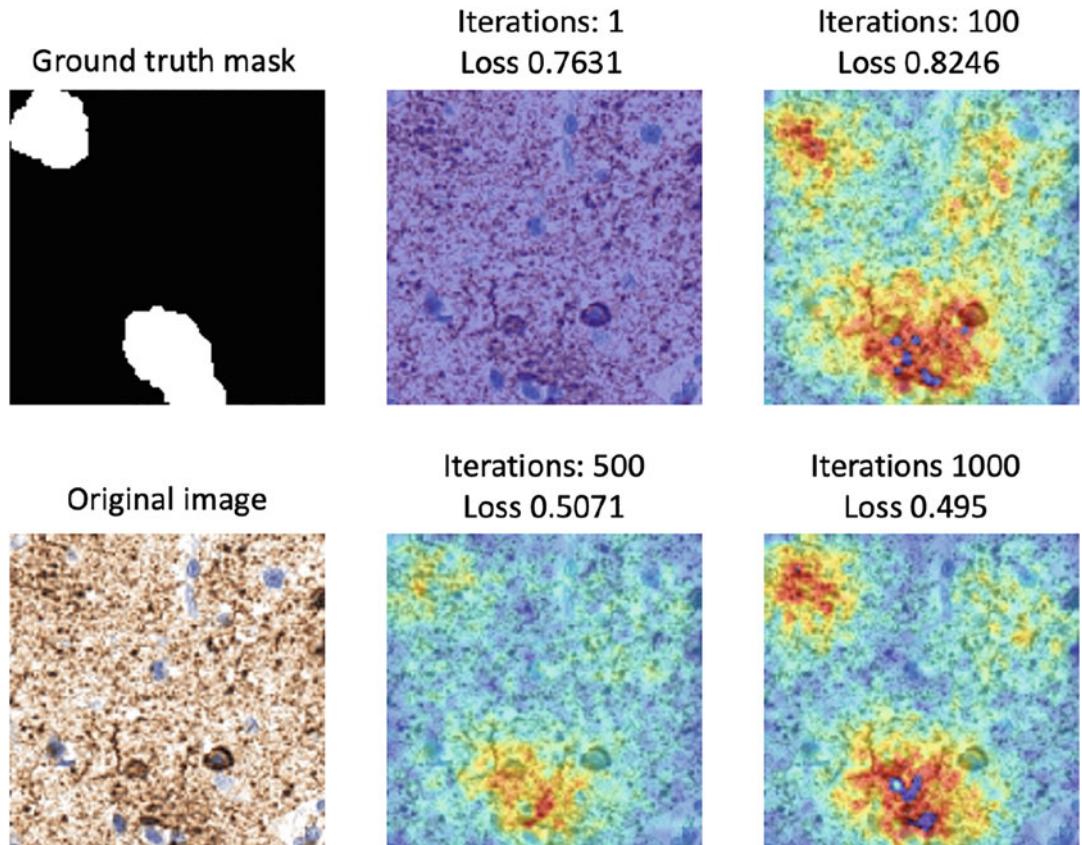


Fig. 10 Attention U-Net results. The figure shows a patch of size 128×128 pixels, the ground-truth binary mask, and the focus progression using successive activation layers of the network

In recent years, explainable AI (xAI) models have been developed to provide insights from and understand the AI decision-making processes by interpreting their second-opinion quantifications, diagnoses, and predictions. Indeed, while explaining simple AI models for regression and classification tasks is relatively straightforward, the explainability task becomes more difficult as the model’s complexity increases. Therefore, a novel paradigm becomes necessary for better interaction between computer scientists, biologists, and clinicians, with the support of an essential new actor: xAI, thus opening the way toward responsible AI: fairness, ethics, privacy, traceability, accountability, safety, and carbon footprint.

In digital histopathology, several studies report on the usage and the benefits of explainable AI models. In [77], the authors describe an xAI-based software named HistoMapr and its application to breast core biopsies. This software automatically identifies the regions of interest (ROI) and rapidly discovers key diagnostic areas from whole slide images of breast cancer biopsies. It generates

a provisional diagnosis based on the automatic detection and classification of relevant ROIs and also provides a list of key findings to pathologists that led to the recommendation. An explainable segmentation pipeline for whole slide images is described in [40], which does a patch-level classification of colon glands for different cancer grades using a CNN followed by inference of class activation maps for the classifier. The activation maps are used for final pixel-level segmentation. The method outperforms other weakly supervised methods applied to these types of images and generalizes to other datasets easily. A medical use-case of AI versus human interpretation of histopathology data using a liver biopsy dataset is described in [78], which also stresses the need to develop methods for causability or measurement of the quality of AI explanations. In [67], AI models like deep auto-encoders were used to generate features from whole-mount prostate cancer pathology images that pathologists could understand. This work showed that a combination of human and AI-generated features produced higher accuracy in predicting prostate cancer recurrence. Finally, in [16], the authors show that, besides providing valuable visual explanation insights, the use of attention U-Net is even helping to increase the results of neuritic plaques segmentation by pulling up the Dice score to 0.75 from 0.72 (with the original U-Net).

Based on the fusion of MRI and histopathology imaging datasets, a deep learning 3D U-Net model with explanations is used in [79] for prostate tumor segmentation. Grad-CAM [80] heat maps were estimated for the last convolutional layer of the U-Net for interpreting the recognition and localization capability of the U-Net. In [81], a framework named NeuroXAI is proposed to render explainability to existing deep learning models in brain imaging research without any architecture modification or reduction in performance. This framework implements seven state-of-the-art explanation methods—including Vanilla gradient [82], Guided back-propagation, Integrated gradients [83], SmoothGrad [84], and Grad-CAM. These methods can be used to generate visual explainability maps for deep learning models like 2D and 3D CNN, VGG [85], and Resnet-50 [86] (for classification) and 2D/3D U-Net (for segmentation). In [87], the high-level features of three deep convolutional neural networks (DenseNet-121, GoogLeNet, MobileNet) are analyzed using the Grad-CAM explainability technique. The Grad-CAM outputs helped distinguish these three models' brain tumor lesion localization capabilities. An explainability framework using SHAP [88] and LIME [89] to predict patient age using the morphological features from a brain MRI dataset is developed in [90]. The SHAP explainability model is robust for this imaging modality to explain morphological feature contributions in predicting age, which would ultimately help develop personalized age-related biomarkers from MRI. Attempts to explain the functional organization of deep segmentation

models like DenseUnet, ResUnet, and SimUnet and understand how these networks achieve high accuracy brain tumor segmentation are presented in [91]. While current xAI methods mainly focus on explaining models on single image modality, the authors of [92] address the explainability issue in multimodal medical images, such as PET-CT or multi-stained pathological images. Combining modality-specific information to explain diagnosis is a complex clinical task, and the authors developed a new multimodal explanation method with modality-specific feature importance.

Intraoperative tissue diagnostic methods have remained unchanged for over 100 years in surgical oncology. Standard light microscopy used in combination with H&E and other staining biomarkers has improved over the last decades with the appearance of new scanner technology. However, the steps involved in the preparation and some artifacts introduced by scanners pose a potential barrier to efficient, reproducible, and accurate intraoperative cancer diagnosis and other brain disorder analyses. As an alternative, label-free optical imaging methods have been developed.

Label-free imaging is a method for cell visualization which does not require labeling or altering the tissue in any way. Bright-field, phase contrast, and differential interference contrast microscopy can be used to visualize label-free cells. The two latter techniques are used to improve the image quality of standard bright-field microscopy. Among its benefits, the cells are analyzed in their unperturbed state, so findings are more reliable and biologically relevant. Also, it is a cheaper and quicker technique as tissue does not need any genetic modification or alteration. In addition, experiments can run longer, making them appropriate for studying cellular dynamics [93]. Raman microscopy, a label-free imaging technique, uses infrared incident light from lasers to capture vibrational signatures of chemical bonds in the tissue sample's molecules. The biomedical tissue is excited with a dual-wavelength fiber laser setup at the so-called *pump* and *Stokes* frequencies to enhance the weak vibrational effect [94]. This technique is known as coherent anti-Stokes Raman scattering (CARS) or stimulated Raman scattering histology (SRH).

Sarri et al. [95] proposed the first one-to-one comparison between SRH and H&E as the latter technique remains the standard in histopathology analyses. The evaluation was conducted using the same cryogenic tissue sample. SRH data was first collected as it did not need staining. SRH and SHG (second harmonic generation, another label-free nonlinear optical technique) were combined to generate a virtual H&E slide for comparison. The results evidenced the almost perfect similarity between SRH and standard H&E slides. Both virtual and real slides show the relevant structures needed to identify cancerous and healthy tissue. In addition, SRH proved to be a fast histologic imaging method suitable for intraoperative procedures.

Similar to standard histopathology, computational methods are also applicable to SRH technology. For instance, Hollon and Orringer [96] proposed a CNN methodology to interpret histologic features from SRH brain tumor images and accurately segment cancerous regions. Results show a slightly better performance (94.6%) than the one obtained by the pathologist (93.9%) in the control group. This study was extended and validated for intraoperative diagnosis in [97]. The study used 2.5 million SRH images and predicted brain tumor diagnosis in under 150 s with an accuracy of 94.6%. The results clearly show the potential of combining computational pathology and stimulated Raman histology for fast and accurate diagnostics in surgical procedures.

Finally, due to its strategic positioning at the cross of molecular biology/omics, radiology/radiomics, and clinics, the rise of computational pathology—by generating “pathomic” features—is expected to play a crucial role in the revolution of spatial transcriptomics, defined as the ability to capture the positional context of transcriptional activity in intact tissue. Spatial transcriptomics is expected to generate a set of technologies allowing researchers to localize transcripts at tissue, cellular, and subcellular levels by providing an unbiased map of RNA molecules in tissue sections. These techniques use microscopy and next-generation sequencing to allow scientists to measure gene expression in a specific tissue or cellular context, consistently paving the road toward more effective personalized medicine. Coupled with these new technologies for data acquisition, we have the release of new WSI brain datasets [98], new frameworks for deep learning analysis of WSI [99, 100], and methods to address the ever-growing concern of privacy and data sharing policies [101].

Acknowledgements

The human samples used in the preparation of the images throughout the chapter were obtained from the Neuro-CEB brain bank⁵ (BRIF Number 0033-00011), partly funded by the patients’ associations ARSEP, ARSLA, “Connaitre les Syndromes Cérébelleux”, France-DFT, France Parkinson and by Vaincre Alzheimer Fondation, to which we express our gratitude. We are also grateful to the patients and their families.

Knowledge and annotations have been provided with the support of Benoît Delatour and Lev Stimmer from the Alzheimer’s Disease team—Paris Brain Institute (ICM), Paris, France.

Some of the research cited here (Subheadings 2, and 4) was supported by Mr Jean-Paul Baudecroux, The Big Brain Theory

⁵ Neuro-CEB brain bank: <https://www.neuroceb.org/en/>.

Program—Paris Brain Institute (ICM) and the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6).

Finally, the authors are grateful to María Gloria Bueno García for reviewing the chapter and providing useful comments.

References

- Serag A, Ion-Margineanu A, Qureshi H, McMillan R, Saint Martin MJ, Diamond J, O'Reilly P, Hamilton P (2019) Translational AI and deep learning in diagnostic pathology. *Front Med* 6:185. <https://doi.org/10.3389/fmed.2019.00185>
- CancerNet Editorial Board (2012) Brain tumor—statistics. <https://www.cancer.net/cancer-types/brain-tumor/statistics>. Accessed 21 May 2022
- Ritchie H (2019) Global mental health: five key insights which emerge from the data. <https://ourworldindata.org/global-mental-health>. Accessed 2 May 2022
- GBD 2017 DALYs and HALE Collaborators (2018) Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 392(10159):1859–1922. [https://doi.org/10.1016/S0140-6736\(18\)32335-3](https://doi.org/10.1016/S0140-6736(18)32335-3)
- European Brain Council (2019) Disease fact sheets. <https://www.braincouncil.eu/disease-fact-sheets/>. Accessed 2 May 2022
- Alzheimer's Association (2022) Alzheimer's and dementia. https://www.alz.org/alzheimer_s_dementia. Accessed 2 May 2022
- GBD 2015 Neurological Disorders Collaborator Group (2017) Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the global burden of disease study 2015. *Lancet Neurol* 16(11):877–897. [https://doi.org/10.1016/S1474-4422\(17\)30299-5](https://doi.org/10.1016/S1474-4422(17)30299-5)
- Stroke Alliance for Europe (2016) About stroke. <https://www.safestroke.eu/about-stroke/>. Accessed 21 May 2022
- Pan American Health Organization (2021) Methodological notes. <https://www.paho.org/en/enlace/technical-notes>. Accessed 2 May 2022
- Peters SR (2009) A practical guide to Frozen section technique, 1st edn. Springer, New York. <https://doi.org/10.1007/978-1-4419-1234-3>
- Yuan Y, Arikath J (2014) Techniques in immunohistochemistry and immunocytochemistry. In: Xiong H, Gendelman HE (eds) Current laboratory methods in neuroscience research. Springer, New York, pp 387–396. https://doi.org/10.1007/978-1-4614-8794-4_27
- Titford M (2009) Progress in the development of microscopical techniques for diagnostic pathology. *J Histotechnol* 32(1):9–19. <https://doi.org/10.1179/his.2009.32.1.9>
- Kapelsohn K, Kapelsohn K (2015) Improved methods for cutting, mounting, and staining tissue for neural histology. Protocol Exchange, Springer Nature. <https://doi.org/10.1038/protex.2015.022>. Protocol (version 1)
- Maňoušková K, Abadie V, Ounissi M, Jimenez G, Stimmer L, Delatour B, Durrelman S, Racoceanu D (2022) Tau protein discrete aggregates in Alzheimer's disease: neuritic plaques and tangles detection and segmentation using computational histopathology. In: Levenson RM, Tomaszewski JE, Ward AD (eds) Medical imaging 2022: digital and computational pathology, SPIE, vol 12039, pp 33–39. <https://doi.org/10.1117/12.2613154>
- Jimenez G, Kar A, Ounissi M, Stimmer L, Delatour B, Racoceanu D (2022) Interpretable deep learning in computational histopathology for refined identification of Alzheimer's disease biomarkers. In: The Alzheimer's Association (ed) Alzheimer's & Dementia: Alzheimer's Association International Conference (AAIC). Wiley, forthcoming
- Jimenez G, Kar A, Ounissi M, Ingrassia L, Boluda S, Delatour B, Stimmer L, Racoceanu D (2022) Visual deep Learning-Based explanation for neuritic plaques segmentation in Alzheimer's disease using weakly annotated whole slide histopathological images. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S

- (eds) Medical image computing and computer assisted intervention (MICCAI). Lecture Notes in Computer Science, vol 13432, Springer Nature Switzerland, pp 336–344. https://doi.org/10.1007/978-3-031-16434-7_33
17. Eiseman E, Bloom G, Brower J, Clancy N, Olmsted SS (2003) Biospecimen collection, processing, annotation, storage, and distribution. In: Case studies of existing human tissue repositories, 1st edn, “Best Practices” for a Biospecimen Resource for the Genomic and Proteomic Era, RAND Corporation, Santa Monica, CA, pp 27–83. <https://doi.org/10.7249/mg120ndc-nci.11>
18. Bolon B, Garman RH, Pardo ID, Jensen K, Sills RC, Roulois A, Radovsky A, Bradley A, Andrews-Jones L, Butt M, Gumprecht L (2013) STP position paper: recommended practices for sampling and processing the nervous system (brain, spinal cord, nerve, and eye) during nonclinical general toxicity studies. *Toxicol Pathol* 41(7):1028–1048. <https://doi.org/10.1177/0192623312474865>
19. Taqi SA, Sami SA, Sami LB, Zaki SA (2018) A review of artifacts in histopathology. *J Oral Maxillofacial Pathol* 22(2):279. https://doi.org/10.4103/jomfp.JOMFP_125_15
20. Jiménez Garay GA (2019) Deep learning for semantic segmentation versus classification in computational pathology: application to mitosis analysis in breast cancer grading. Master’s thesis, Pontificia Universidad Católica del Perú
21. Garcia Rojo M, Bueno García G, Pece Mateos C, González García J, Carboj Vicente M (2006) Critical comparison of 31 commercially available digital slide systems in pathology. *Int J Surg Pathol* 14(4):285–305. <https://doi.org/10.1177/106689690692274>
22. Farahani N, Parwani AV, Pantanowitz L (2015) Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathol Lab Med Int* 7:23–33. <https://doi.org/10.2147/PLMI.S59826>
23. DICOM Standards Committee P Working Groups 26 (2010) Supplement 145: whole slide microscopic image IOD and SOP classes. Tech. Rep. 145. National Electrical Manufacturers Association (NEMA), Virginia, United States
24. Lajara N, Espinosa-Aranda JL, Deniz O, Bueno G (2019) Optimum web viewer application for DICOM whole slide image visualization in anatomical pathology. *Comput Methods Progr Biomed* 179:104983. <https://doi.org/10.1016/j.cmpb.2019.104983>
25. Helin H, Tolonen T, Ylinen O, Tolonen P, Näpänkangas J, Isola J (2018) Optimized JPEG 2000 compression for efficient storage of histopathological whole-slide images. *J Pathol Inform* 9:20. https://doi.org/10.4103/jpi.jpi_69_17
26. Bauer TW, Slaw RJ, McKenney JK, Patil DT (2015) Validation of whole slide imaging for frozen section diagnosis in surgical pathology. *J Pathol Inform* 6(1):49. <https://doi.org/10.4103/2153-3539.163988>
27. Kong J, Cooper LAD, Wang F, Gao J, Teodoro G, Scarpace L, Mikkelsen T, Schniederjan MJ, Moreno CS, Saltz JH, Brat DJ (2013) Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates. *Plos One* 8(11):e81049. <https://doi.org/10.1371/journal.pone.0081049>
28. Lu H, Papathomas TG, van Zessen D, Palli I, de Krijger RR, van der Spek PJ, Dinjens WNM, Stubbs AP (2014) Automated selection of hotspots (ASH): enhanced automated segmentation and adaptive step finding for Ki67 hotspot detection in adrenal cortical cancer. *Diagn Pathol* 9:216. <https://doi.org/10.1186/s13000-014-0216-6>
29. Yeh FC, Parwani AV, Pantanowitz L, Ho C (2014) Automated grading of renal cell carcinoma using whole slide imaging. *J Pathol Inform* 5(1):23. <https://doi.org/10.4103/2153-3539.137726>
30. Pantanowitz L, Sinard JH, Henricks WH, Fatheree LA, Carter AB, Contis L, Beckwith BA, Evans AJ, Lal A, Parwani AV, College of American Pathologists Pathology and Laboratory Quality Center (2013) Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med* 137(12):1710–1722. <https://doi.org/10.5858/arpa.2013-0093-CP>
31. Office of the Commissioner-FDA (2017) Press Announcements—FDA allows marketing of first whole slide imaging system for digital pathology. <https://www.fda.gov/news-events/newsroom/pressannouncements/ucm552742.htm>. Accessed 20 Dec 2021
32. Dimitriou N, Arandjelović O, Caie PD (2019) Deep learning for whole slide image analysis: an overview. *Front Med (Lausanne)* 6:264. <https://doi.org/10.3389/fmed.2019.00264>

33. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, Schmitt C, Thomas NE (2009) A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE international symposium on biomedical imaging (ISBI): from nano to macro, IEEE, pp 1107–1110. <https://doi.org/10.1109/ISBI.2009.5193250>
34. Vahadane A, Peng T, Albarqouni S, Baust M, Steiger K, Schlitter AM, Sethi A, Esposito I, Navab N (2015) Structure-preserved color normalization for histological images. In: 2015 IEEE 12th international symposium on biomedical imaging (ISBI). IEEE, pp 1012–1015. <https://doi.org/10.1109/ISBI.2015.7164042>
35. Reinhard E, Adhikhmin M, Gooch B, Shirley P (2001) Color transfer between images. *IEEE Comput Graph Appl* 21(5):34–41. <https://doi.org/10.1109/38.946629>
36. Magee D, Treanor D, Crellin D, Shires M, Mohee K, Quirke P (2009) Colour normalisation in digital histopathology images. In: Elson D, Rajpoot N (eds) Optical tissue image analysis in microscopy, histopathology and endoscopy: OPTIMHisE, MICCAI workshop, pp 100–111
37. Kang H, Luo D, Feng W, Zeng S, Quan T, Hu J, Liu X (2021) StainNet: a fast and robust stain normalization network. *Front Med* 8: 746307. <https://doi.org/10.3389/fmed.2021.746307>
38. Runz M, Rusche D, Schmidt S, Weihrauch MR, Hesser J, Weis CA (2021) Normalization of HE-stained histological images using cycle consistent generative adversarial networks. *Diagn Pathol* 16(1):71. <https://doi.org/10.1186/s13000-021-01126-y>
39. Jiménez G, Racoceanu D (2019) Deep learning for semantic segmentation vs. classification in computational pathology: application to mitosis analysis in breast cancer grading. *Front Bioeng Biotechnol* 7:145. <https://doi.org/10.3389/fbioe.2019.00145>
40. Chan L, Hosseini M, Rowsell C, Plataniotis K, Damaskinos S (2019) HistosegNet: semantic segmentation of histological tissue type in whole slide images. In: 2019 IEEE/CVF international conference on computer vision (ICCV). IEEE, pp 10661–10670. <https://doi.org/10.1109/ICCV.2019.01076>
41. Ahmedt-Aristizabal D, Armin MA, Denman S, Fookes C, Petersson L (2022) A survey on graph-based deep learning for computational histopathology. *Comput Med Imaging Graph* 95:102027. <https://doi.org/10.1016/j.compmedimag.2021.102027>
42. Anklin V, Pati P, Jaume G, Bozorgtabar B, Foncubierta-Rodriguez A, Thiran JP, Sibony M, Gabrani M, Goksel O (2021) Learning whole-slide segmentation from inexact and incomplete labels using tissue graphs. In: de Bruijne M, Cattin PC, Cotin S, Padov N, Speidel S, Zheng Y, Essert C (eds) Medical image computing and computer assisted intervention (MICCAI). Lecture notes in computer science, vol 12902. Springer, Berlin, pp 636–646. https://doi.org/10.1007/978-3-030-87196-3_59
43. Tellez D, Litjens G, van der Laak J, Ciompi F (2021) Neural image compression for gigapixel histopathology image analysis. *IEEE Trans Pattern Anal Mach Intell* 43(2): 567–578. <https://doi.org/10.1109/TPAMI.2019.2936841>
44. Zhu X, Yao J, Zhu F, Huang J (2017) WSISA: making survival prediction from whole slide histopathological images. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 6855–6863. <https://doi.org/10.1109/CVPR.2017.725>
45. Qaiser T, Rajpoot NM (2019) Learning where to see: a novel attention model for automated immunohistochemical scoring. *IEEE Trans Med Imaging* 38(11): 2620–2631. <https://doi.org/10.1109/TMI.2019.2907049>
46. Momeni A, Thibault M, Gevaert O (2018) Deep recurrent attention models for histopathological image analysis. bioRxiv preprint. <https://doi.org/10.1101/438341>
47. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH (2016) Patch-based convolutional neural network for whole slide tissue image classification. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2424–2433. <https://doi.org/10.1109/CVPR.2016.266>
48. Campanella G, Silva VWK, Fuchs TJ (2018) Terabyte-scale deep multiple instance learning for classification and localization in pathology. arXiv preprint. <https://doi.org/10.48550/ARXIV.1805.06983>
49. Liu Y, Gadepalli K, Norouzi M, Dahl GE, Kohlberger T, Boyko A, Venugopalan S, Timofeev A, Nelson PQ, Corrado G, Hipp J, Peng L, Stumpe MC (2017) Detecting cancer metastases on gigapixel pathology images. arXiv preprint. <https://doi.org/10.48550/ARXIV.1703.02442>
50. van Rijthoven M, Balkenhol M, Siliqa K, van der Laak J, Ciompi F (2021) HookNet: multi-

- resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Med Image Anal* 68: 101890. <https://doi.org/10.1016/j.media.2020.101890>
51. Schmitz R, Madesta F, Nielsen M, Krause J, Steurer S, Werner R, Rösch T (2021) Multi-scale fully convolutional neural networks for histopathology image segmentation: from nuclear aberrations to the global tissue architecture. *Med Image Anal* 70:101996. <https://doi.org/10.1016/j.media.2021.101996>
 52. Janowczyk A, Madabhushi A (2016) Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Inform* 7:29. <https://doi.org/10.4103/2153-3539.186902>
 53. Kong J, Cooper L, Wang F, Chisolm C, Moreno C, Kurc T, Widener P, Brat D, Saltz J (2011) A comprehensive framework for classification of nuclei in digital microscopy imaging: an application to diffuse gliomas. In: 2011 IEEE international symposium on biomedical imaging (ISBI): from nano to macro. IEEE, pp 2128–2131. <https://doi.org/10.1109/ISBI.2011.5872833>
 54. Xing F, Xie Y, Yang L (2016) An automatic learning-based framework for robust nucleus segmentation. *IEEE Trans Med Imaging* 35(2):550–566. <https://doi.org/10.1109/TMI.2015.2481436>
 55. Xu Y, Jia Z, Ai Y, Zhang F, Lai M, Chang EIC (2015) Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 947–951. <https://doi.org/10.1109/ICASSP.2015.7178109>
 56. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems (NIPS). Curran Associates, vol 25
 57. Xu Y, Jia Z, Wang LB, Ai Y, Zhang F, Lai M, Chang EIC (2017) Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinform* 18(1):281. <https://doi.org/10.1186/s12859-017-1685-x>
 58. Ker J, Bai Y, Lee HY, Rao J, Wang L (2019) Automated brain histology classification using machine learning. *J Clin Neurosci* 66:239–245. <https://doi.org/10.1016/j.jocn.2019.05.019>
 59. Truong AH, Sharmanska V, Limbäck-Stanic, Grech-Sollars M (2020) Optimization of deep learning methods for visualization of tumor heterogeneity and brain tumor grading through digital pathology. *Neuro-Oncol Adv* 2(1):vdaa110. <https://doi.org/10.1093/noajnl/vdaa110>
 60. Zadeh Shirazi A, Fornaciari E, McDonnell MD, Yaghoobi M, Cevallos Y, Tello-Oquendo L, Inca D, Gomez GA (2020) The application of deep convolutional neural networks to brain cancer images: a survey. *J Pers Med* 10(4):224. <https://doi.org/10.3390/jpm10040224>
 61. Wurts A, Oakley DH, Hyman BT, Samsi S (2020) Segmentation of Tau stained Alzheimer's brain tissue using convolutional neural networks. In: 2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC). IEEE, vol 2020, pp 1420–1423. <https://doi.org/10.1109/EMBC44109.2020.9175832>
 62. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A (eds) Medical image computing and computer-assisted intervention (MICCAI). Lecture notes in computer science. Springer, Berlin, pp 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
 63. Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell* 39(12):2481–2495. <https://doi.org/10.1109/TPAMI.2016.2644615>
 64. Signaevsky M, Prastawa M, Farrell K, Tabish N, Baldwin E, Han N, Iida MA, Koll J, Bryce C, Purohit D, Haroutunian V, McKee AC, Stein TD, White CL 3rd, Walker J, Richardson TE, Hanson R, Donovan MJ, Cordon-Cardo C, Zeineh J, Fernandez G, Crary JF (2019) Artificial intelligence in neuropathology: deep learning-based assessment of tauopathy. *Lab Invest* 99(7):1019–1029. <https://doi.org/10.1038/s41374-019-0202-4>
 65. Vega AR, Chkhcheidze R, Jarmale V, Shang P, Foong C, Diamond MI, White CL 3rd, Rajaram S (2021) Deep learning reveals disease-specific signatures of white matter pathology in tauopathies. *Acta neuropathologica communications* 9(1):170. <https://doi.org/10.1186/s40478-021-01271-x>
 66. Border SP, Sarder P (2021) From what to why, the growing need for a focus shift toward explainability of AI in digital pathology. *Front*

- Physiol 12:821217. <https://doi.org/10.3389/fphys.2021.821217>
67. Yamamoto Y, Tsuzuki T, Akatsuka J, Ueki M, Morikawa H, Numata Y, Takahara T, Tsuyuki T, Tsutsumi K, Nakazawa R, Shimizu A, Maeda I, Tsuchiya S, Kanno H, Kondo Y, Fukumoto M, Tamiya G, Ueda N, Kimura G (2019) Automated acquisition of explainable knowledge from unannotated histopathology images. Nat Commun 10(1):5642. <https://doi.org/10.1038/s41467-019-13647-8>
 68. Lahiani A (2020) Deep learning solutions for cancer drug development in digital pathology. PhD thesis, Technische Universität München
 69. Kurc T, Bakas S, Ren X, Bagari A, Momeni A, Huang Y, Zhang L, Kumar A, Thibault M, Qi Q, Wang Q, Kori A, Gevaert O, Zhang Y, Shen D, Khened M, Ding X, Krishnamurthi G, Kalpathy-Cramer J, Davis J, Zhao T, Gupta R, Saltz J, Farahani K (2020) Segmentation and classification in digital pathology for glioma research: challenges and deep learning approaches. Front Neurosci 14:27. <https://doi.org/10.3389/fnins.2020.00027>
 70. Qi Q, Li Y, Wang J, Zheng H, Huang Y, Ding X, Rohde GK (2019) Label-efficient breast cancer histopathological image classification. IEEE J Biomed Health Inform 23(5):2108–2116. <https://doi.org/10.1109/JBHI.2018.2885134>
 71. Wang X, Wang R, Yang S, Zhang J, Wang M, Zhong D, Zhang J, Han X (2022) Combining radiology and pathology for automatic glioma classification. Front Bioeng Biotechnol 10:841958. <https://doi.org/10.3389/fbioe.2022.841958>
 72. Hsu WW, Guo JM, Pei L, Chiang LA, Li YF, Hsiao JC, Colen R, Liu P (2022) A weakly supervised deep learning-based method for glioma subtype classification using WSI and mpMRIs. Sci Rep 12(1):6111. <https://doi.org/10.1038/s41598-022-09985-1>
 73. Duyckaerts C, Delatour B, Potier MC (2009) Classification and basic pathology of Alzheimer disease. Acta Neuropathol 118(1):5–36. <https://doi.org/10.1007/s00401-009-0532-1>
 74. Koga S, Ikeda A, Dickson DW (2022) Deep learning-based model for diagnosing Alzheimer's disease and tauopathies. Neuropathol Appl Neurobiol 48(1):e12759. <https://doi.org/10.1111/nan.12759>
 75. Ushizima D, Chen Y, Alegra M, Ovando D, Eser R, Lee W, Poon K, Shankar A, Kantamneni N, Satrawada S, Junior EA, Heinsen H, Tosun D, Grinberg LT (2022) Deep learning for Alzheimer's disease: mapping large-scale histological tau protein for neuroimaging biomarker validation. NeuroImage 248:118790. <https://doi.org/10.1016/j.neuroimage.2021.118790>
 76. Jonkman LE, Kenkhuis B, Geurts JJG, van de Berg WDJ (2019) Post-mortem MRI and histopathology in neurologic disease: a translational approach. Neurosc Bull 35(2):229–243. <https://doi.org/10.1007/s12264-019-00342-3>
 77. Tosun AB, Pullara F, Becich MJ, Taylor DL, Fine JL, Chennubhotla SC (2020) Explainable AI (xAI) for anatomic pathology. Adv Anat Pathol 27(4):241–250. <https://doi.org/10.1097/PAP.0000000000000264>
 78. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H (2019) Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Mining Knowl Discovery 9(4):e1312. <https://doi.org/10.1002/widm.1312>
 79. Gunashekar DD, Bielak L, Häggle L, Oerther B, Benndorf M, Grosu AL, Brox T, Zamboglou C, Bock M (2022) Explainable AI for CNN-based prostate tumor segmentation in multi-parametric MRI correlated to whole mount histopathology. Radiat Oncol 17(1):65. <https://doi.org/10.1186/s13014-022-02035-0>
 80. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2020) Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 128(2):336–359. <https://doi.org/10.1007/s11263-019-01228-7>
 81. Zeineldin RA, Karar ME, Elshaer Z, Coburger J, Wirtz CR, Burgert O, Mathis-Ullrich F (2022) Explainability of deep neural networks for MRI analysis of brain tumors. Int J Comput Assisted Radiol Surg 17(9):1673–1683. <https://doi.org/10.1007/s11548-022-02619-x>
 82. Simonyan K, Vedaldi A, Zisserman A (2014) Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint. <https://doi.org/10.48550/ARXIV.1312.6034>
 83. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning (ICML), PMLR, Proceedings of machine learning research, vol 70, pp 3319–3328
 84. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M (2017) SmoothGrad: removing noise by adding noise. arXiv preprint.

- <https://doi.org/10.48550/ARXIV.1706.03825>
85. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint. <https://doi.org/10.48550/arXiv.1409.1556>
86. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 770–778. <https://doi.org/10.1109/CVPR.2016.90>
87. Esmaeili M, Vettukattil R, Banitalebi H, Krogh NR, Geitung JT (2021) Explainable artificial intelligence for human-machine interaction in brain tumor localization. *J Pers Med* 11(11). <https://doi.org/10.3390/jpm1111213>
88. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Proceedings of the 31st international conference on neural information processing systems (NIPS). Curran Associates, vol 30, pp 4768–4777
89. Ribeiro M, Singh S, Guestrin C (2016) “why should I trust you?” Explaining the predictions of any classifier. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: demonstrations (NAACL). Association for Computational Linguistics, pp 97–101. <https://doi.org/10.18653/v1/n1-6-3020>
90. Lombardi A, Diacono D, Amoroso N, Monaco A, Tavares JMRS, Bellotti R, Tangaro S (2021) Explainable deep learning for personalized age prediction with brain morphology. *Front Neurosci* 15:674055. <https://doi.org/10.3389/fnins.2021.674055>
91. Natekar P, Kori A, Krishnamurthi G (2020) Demystifying brain tumor segmentation networks: interpretability and uncertainty analysis. *Front Comput Neurosci* 14:6. <https://doi.org/10.3389/fncom.2020.00006>
92. Jin W, Li X, Hamarneh G (2022) Evaluating explainable AI on a Multi-modal medical imaging task: can existing algorithms fulfill clinical requirements? In: Proceedings of the AAAI conference on artificial intelligence (AAAI), vol 36. AAAI Press, pp 11945–11953. <https://doi.org/10.1609/aaai.v36i11.21452>
93. Bleloch J (2020) Label-free imaging of live cells. <https://cytoshare.com/resources/resources/label-free-imaging-live-cells>. Accessed 25 May 2022
94. Marx V (2019) It’s free imaging—label-free, that is. *Nature Methods* 16(12):1209–1212. <https://doi.org/10.1038/s41592-019-0664-8>
95. Sarri B, Poizat F, Heuke S, Wojak J, Franchi F, Caillol F, Giovannini M, Rigneault H (2019) Stimulated raman histology: one to one comparison with standard hematoxylin and eosin staining. *Biomed Opt Express* 10(10): 5378–5384. <https://doi.org/10.1364/BOE.10.0005378>
96. Hollon TC, Orringer DA (2020) An automated tissue-to-diagnosis pipeline using intraoperative stimulated raman histology and deep learning. *Mol Cell Oncol* 7(3): 1736742. <https://doi.org/10.1080/23723556.2020.1736742>
97. Hollon TC, Pandian B, Adapa AR, Urias E, Save AV, Khalsa SSS, Eichberg DG, D’Amico RS, Farooq ZU, Lewis S, Petridis PD, Marie T, Shah AH, Garton HJL, Maher CO, Heth JA, McKean EL, Sullivan SE, Hervey-Jumper SL, Patil PG, Thompson BG, Sagher O, McKhann GM 2nd, Komotar RJ, Ivan ME, Snuderl M, Otten ML, Johnson TD, Sisti MB, Bruce JN, Muraszko KM, Trautman J, Freudiger CW, Canoll P, Lee H, Camelo-Piragua S, Orringer DA (2020) Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat Med* 26(1):52–58. <https://doi.org/10.1038/s41591-019-0715-9>
98. Roetzer-Pejrimovsky T, Moser AC, Atli B, Vogel CC, Mercea PA, Prihoda R, Gelpi E, Haberler C, Höftberger R, Hainfellner JA, Baumann B, Langs G, Woehrer A (2022) The digital brain tumour atlas, an open histopathology resource. *Sci Data* 9(1):55. <https://doi.org/10.1038/s41597-022-01157-0>
99. Berman AG, Orchard WR, Gehring M, Mar-kowitz F (2021) PathML: a unified framework for whole-slide image analysis with deep learning. medRxiv preprint. <https://doi.org/10.1101/2021.07.07.21260138>
100. Hägele M, Seegerer P, Lapuschkin S, Bockmayr M, Samek W, Klauschen F, Müller KR, Binder A (2020) Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Sci Rep* 10(1):6423. <https://doi.org/10.1038/s41598-020-62724-2>
101. Lu MY, Chen RJ, Kong D, Lipkova J, Singh R, Williamson DFK, Chen TY, Mahmood F (2022) Federated learning for computational pathology on gigapixel whole slide images. *Med Image Anal* 76:102298.

- <https://doi.org/10.1016/j.media.2021.102298>
102. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B, Rueckert D (2018) Attention U-Net: learning where to look for the pancreas. arXiv preprint. <https://doi.org/10.48550/ARXIV.1804.03999>
 103. Roy AG, Siddiqui S, Pölsterl S, Navab N, Wachinger C (2019) BrainTorrent: a peer-to-peer environment for decentralized federated learning. arXiv preprint. <https://doi.org/10.48550/ARXIV.1905.06731>
 104. Spotorno N, Coughlin DG, Olm CA, Wolk D, Vaishnavi SN, Shaw LM, Dahodwala N, Morley JF, Duda JE, Deik AF, Spindler MA, Chen-Plotkin A, Lee EB, Trojanowski JQ, McMillan CT, Weintraub D, Grossman M, Irwin DJ (2020) Tau pathology associates with in vivo cortical thinning in Lewy body disorders. Ann Clin Transl Neurol 7(12):2342–2355. <https://doi.org/10.1002/acn3.51183>
 105. Banerji S, Mitra S (2022) Deep learning in histopathology: a review. WIREs Data Mining Knowl Discovery 12(1):e1439. <https://doi.org/10.1002/widm.1439>
 106. Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S (2019) Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. In: Crimi A, Bakas S, Kuifj H, Keyvan F, Reyes M, van Walsum T (eds) Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries (BrainLes). Lecture notes in computer science, vol 11383. Springer, Berlin, pp 92–104. https://doi.org/10.1007/978-3-030-11723-8_9
 107. Li J, Chen W, Huang X, Yang S, Hu Z, Duan Q, Metaxas DN, Li H, Zhang S (2021) Hybrid supervision learning for pathology whole slide image classification. In: de Bruijne M, Cattin PC, Cotin S, Padov N, Speidel S, Zheng Y, Essert C (eds) Medical image computing and computer assisted intervention (MICCAI). Lecture notes in computer science, vol 12908. Springer, Berlin, pp 309–318. https://doi.org/10.1007/978-3-030-87237-3_30
 108. Barker J, Hoogi A, Depeursinge A, Rubin DL (2016) Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. Med Image Anal 30:60–71. <https://doi.org/10.1016/j.media.2015.12.002>
 109. Madabhushi A, Lee G (2016) Image analysis and machine learning in digital pathology: challenges and opportunities. Med Image Anal 33:170–175. <https://doi.org/10.1016/j.media.2016.06.037>
 110. Srinidhi CL, Ciga O, Martel AL (2021) Deep neural network models for computational histopathology: a survey. Med Image Anal 67: 101813. <https://doi.org/10.1016/j.media.2020.101813>
 111. Kasanuki K, Ferman TJ, Murray ME, Heckman MG, Pedraza O, Hanna Al-Shaikh FS, Mishima T, Diehl NN, van Gerpen JA, Uitti RJ, Wszolek ZK, Graff-Radford NR, Dickson DW (2018) Daytime sleepiness in dementia with Lewy bodies is associated with neuronal depletion of the nucleus basalis of Meynert. Parkinsonism Relat Disord 50:99–103. <https://doi.org/10.1016/j.parkreldis.2018.02.003>
 112. Seeley EH, Caprioli RM (2011) MALDI imaging mass spectrometry of human tissue: method challenges and clinical perspectives. Trends Biotechnol 29(3):136–143. <https://doi.org/10.1016/j.tibtech.2010.12.002>
 113. Blow N (2007) Tissue preparation: tissue issues. Nature 448(7156):959–963. <https://doi.org/10.1038/448959a>
 114. Kim H, Yoon H, Thakur N, Hwang G, Lee EJ, Kim C, Chong Y (2021) Deep learning-based histopathological segmentation for whole slide images of colorectal cancer in a compressed domain. Sci Rep 11(1):22520. <https://doi.org/10.1038/s41598-021-01905-z>
 115. Khened M, Kori A, Rajkumar H, Krishnamurthi G, Srinivasan B (2021) A generalized deep learning framework for whole-slide image segmentation and analysis. Sci Rep 11(1):11579. <https://doi.org/10.1038/s41598-021-90444-8>
 116. Shakir MN, Dugger BN (2022) Advances in deep neuropathological phenotyping of Alzheimer disease: past, present, and future. Journal of neuropathology and experimental neurology 81(1):2–15. <https://doi.org/10.1093/jnen/nlab122>
 117. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, Hawkins C, Ng HK, Pfister SM, Reifenberger G, Soffietti R, von Deimling A, Ellison DW (2021) The 2021 WHO classification of tumors of the central nervous system: a summary. Neuro-Oncology 23(8):1231–1251. <https://doi.org/10.1093/neuonc/noab106>
 118. Vandenberghe ME, Balbastre Y, Souedet N, Hérard AS, Dhenain M, Frouin F, Delzescaux T (2015) Robust supervised segmentation of neuropathology whole-slide microscopy images. In: 2015 37th annual international conference of the IEEE engineering in

- medicine and biology society (EMBC). IEEE, pp 3851–3854. <https://doi.org/10.1109/EMBC.2015.7319234>
119. Bández P, van de Loo R, Intezar M, Geijs D, Ciompi F, van Ginneken B, van der Laak J, Litjens G (2017) Comparison of different methods for tissue segmentation in histopathological whole-slide images. In: 2017 IEEE 14th international symposium on biomedical imaging (ISBI), pp 591–595. <https://doi.org/10.1109/ISBI.2017.7950590>
120. Xiao Y, Decencière E, Velasco-Forero S, Burdin H, Bornschlögl T, Bernerd F, Warrick E, Baldeweg T (2019) A new color augmentation method for deep learning segmentation of histological images. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI), pp 886–890. <https://doi.org/10.1109/ISBI.2019.8759591>
121. Li Y, Xu Z, Wang Y, Zhou H, Zhang Q (2020) SU-Net and DU-Net fusion for tumour segmentation in histopathology images. In: 2020 IEEE 17th international symposium on biomedical imaging (ISBI). IEEE, pp 461–465. <https://doi.org/10.1109/ISBI45749.2020.9098678>
122. Pan SJ, Yang Q (2010) A survey on transfer learning. *IEEE Trans Knowl Data Eng* 22(10):1345–1359. <https://doi.org/10.1109/TKDE.2009.191>
123. Caie PD, Schuur K, Oniscu A, Mullen P, Reynolds PA, Harrison DJ (2013) Human tissue in systems medicine. *FEBS J* 280(23): 5949–5956. <https://doi.org/10.1111/febs.12550>
124. Snead DRJ, Tsang YW, Meskini A, Kimani PK, Crossman R, Rajpoot NM, Blessing E, Chen K, Gopalakrishnan K, Matthews P, Momtahan N, Read-Jones S, Sah S, Simmons E, Sinha B, Suortamo S, Yeo Y, El Daly H, Cree IA (2016) Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology* 68(7): 1063–1072. <https://doi.org/10.1111/his.12879>
125. Williams JH, Mepham BL, Wright DH (1997) Tissue preparation for immunocytochemistry. *J Clin Pathol* 50(5):422–428. <https://doi.org/10.1136/jcp.50.5.422>
126. Qin P, Chen J, Zeng J, Chai R, Wang L (2018) Large-scale tissue histopathology image segmentation based on feature pyramid. *EURASIP J Image Video Process* 2018(1):1–9. <https://doi.org/10.1186/s13640-018-0320-8>
127. Wong DR, Tang Z, Mew NC, Das S, Athey J, McAleese KE, Kofler JK, Flanagan ME, Borys E, White CL 3rd, Butte AJ, Dugger BN, Keiser MJ (2022) Deep learning from multiple experts improves identification of amyloid neuropathologies. *Acta Neuropathol Commun* 10(1):66. <https://doi.org/10.1186/s40478-022-01365-0>
128. Willuweit A, Velden J, Godemann R, Manook A, Jetzek F, Tintrup H, Kauselmann G, Zevnik B, Henriksen G, Drzezga A, Pohlner J, Schoor M, Kemp JA, von der Kammer H (2009) Early-onset and robust amyloid pathology in a new homozygous mouse model of Alzheimer's disease. *Plos One* 4(11):e7931. <https://doi.org/10.1371/journal.pone.0007931>
129. Khalsa SSS, Hollon TC, Adapa A, Urias E, Srinivasan S, Jairath N, Szczepanski J, Ouillette P, Camelo-Piragua S, Orringer DA (2020) Automated histologic diagnosis of CNS tumors with machine learning. *CNS Oncol* 9(2):CNS56. <https://doi.org/10.2217/cns-2020-0003>
130. Tandel GS, Biswas M, Kakde OG, Tiwari A, Suri HS, Turk M, Laird JR, Asare CK, Ankrah AA, Khanna NN, Madhusudhan BK, Saba L, Suri JS (2019) A review on a deep learning perspective in brain cancer classification. *Cancers (Basel)* 11(1):111. <https://doi.org/10.3390/cancers11010111>
131. Jose L, Liu S, Russo C, Nadort A, Di Ieva A (2021) Generative adversarial networks in digital pathology and histopathological image processing: a review. *J Pathol Inform* 12:43. https://doi.org/10.4103/jpi.jpi_103_20
132. Herrmann MD, Clunie DA, Fedorov A, Doyle SW, Pieper S, Klepeis V, Le LP, Mutter GL, Milstone DS, Schultz TJ, Kikinis R, Kotecha GK, Hwang DH, Andriole KP, Iafurrat AJ, Brink JA, Boland GW, Dreyer KJ, Michalski M, Golden JA, Louis DN, Lennerz JK (2018) Implementing the DICOM standard for digital pathology. *J Pathol Inform* 9: 37. https://doi.org/10.4103/jpi.jpi_42_18
133. Aeffner F, Zarella MD, Buchbinder N, Bui MM, Goodman MR, Hartman DJ, Lujan GM, Molani MA, Parwani AV, Lillard K, Turner OC, Vemuri VNP, Yuil-Valdes AG, Bowman D (2019) Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. *J Pathol Inform* 10:9. https://doi.org/10.4103/jpi.jpi_82_18
134. Farahani K, Kurc T, Bakas S, Bearce BA, Kalpathy-Cramer J, Freymann J, Saltz J,

- Stahlberg E, Zaki G, Nasrallah MP, Shinohara RT (2020) Computational precision medicine radiology-pathology challenge on brain tumor classification 2020. In: 23rd International conference on medical image computing and computer assisted intervention (MICCAI challenge). <https://doi.org/10.5281/ZENODO.3718894>
135. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Bach F, Blei D (eds) Proceedings of the 32nd international conference on machine learning (ICML), PMLR, Proceedings of machine learning research, vol 37, pp 448–456
136. DICOM Standards Committee P Working Groups 26 (2008) Supplement 122: specimen module and revised pathology SOP classes. Tech. Rep. 122, National Electrical Manufacturers Association (NEMA), Virginia, United States
137. Wright JR Jr (1985) The development of the frozen section technique, the evolution of surgical biopsy, and the origins of surgical pathology. Bull Hist Med 59(3):295–326

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made. The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 19

Integration of Multimodal Data

Marco Lorenzi, Marie Deprez, Irene Balelli, Ana L. Aguila, and Andre Altmann

Abstract

This chapter focuses on the joint modeling of heterogeneous information, such as imaging, clinical, and biological data. This kind of problem requires to generalize classical uni- and multivariate association models to account for complex data structure and interactions, as well as high data dimensionality.

Typical approaches are essentially based on the identification of latent modes of maximal statistical association between different sets of features and ultimately allow to identify joint patterns of variations between different data modalities, as well as to predict a target modality conditioned on the available ones. This rationale can be extended to account for several data modalities jointly, to define multi-view, or multi-channel, representation of multiple modalities. This chapter covers both classical approaches such as partial least squares (PLS) and canonical correlation analysis (CCA), along with most recent advances based on multi-channel variational autoencoders. Specific attention is here devoted to the problem of interpretability and generalization of such high-dimensional models. These methods are illustrated in different medical imaging applications, and in the joint analysis of imaging and non-imaging information, such as -omics or clinical data.

Key words Multivariate analysis, Latent variable models, Multimodal imaging, -Omics, Imaging-genetics, Partial least squares, Canonical correlation analysis, Variational autoencoders, Sparsity, Interpretability

1 Introduction

The goal of multimodal data analysis is to reveal novel insights on complex biological conditions. Through the combined analysis of multiple type of data, and the complementary views on pathophysiological processes they provide, we have the potential to improve our understanding of the underlying processes leading to complex and multifactorial disorders [1]. In medical imaging applications, multiple imaging modalities, such as structural magnetic resonance imaging (sMRI), functional MRI (fMRI), diffusion tensor imaging (DTI), or positron emission tomography (PET), can be jointly analyzed to better characterize pathological conditions affecting individuals [2]. Other typical multimodal analysis problems involve

the joint analysis of heterogeneous data types, such as imaging and genetics data, where medical imaging is associated with the patient's genotype information, represented by genetic variants such as single-nucleotide polymorphisms (SNPs) [3]. This kind of application, termed *imaging-genetics*, is of central importance for the identification of genetic risk factors underlying complex diseases including age-related macular degeneration, obesity, schizophrenia, and Alzheimer's disease [4].

Despite the great potential of multimodal data analysis, the complexity of multiple data types and clinical questions poses several challenges to the researchers, involving scalability, interpretability, and generalization of complex association models.

1.1 Challenges of Multimodal Data Assimilation

Due to the complementary nature of multimodal information, there is great interest in combining different data types to better characterize the anatomy and physiology of patients and individuals. Multimodal data is generally acquired using heterogeneous protocols highlighting different anatomical, physiological, clinical, and biological information for a given individual [5].

Typical multimodal data integration challenges are:

- *Non-commensurability.* Since each data modality quantifies different physical and biological phenomena, multimodal data is represented by heterogeneous physical units associated to different aspects of the studied biological process (e.g., brain structure, activity, clinical scores, gene expression levels).
- *Spatial heterogeneity.* Multimodal medical images are characterized by specific spatial resolution, which is independent from the spatial coordinate system on which they are standardized.
- *Heterogeneous dimensions.* The data type and dimensions of medical data can vary according to the modality, ranging from scalars and time series typical of fMRI and PET data to structured tensors of diffusion weighted imaging.
- *Heterogeneous noise.* Medical data modalities are characterized by specific and heterogeneous artifacts and measurement uncertainty, resulting from heterogeneous acquisition and processing routines.
- *Missing data.* Multimodal medical datasets are often incomplete, since patients may not undergo the same protocol, and some modalities may be more expensive to acquire than others.
- *Interpretability.* A major challenge of multimodal data integration is the interpretability of the analysis results. This aspect is impacted by the complexity of the analysis methods and generally requires important expertise in data acquisition, processing, and analysis.

Multimodal data analysis methods proposed in the literature have been focusing on different data complexity and integration, depending on the application of interest. Visual inspection is the typical initial step of multimodal studies, where single modalities are compared on a qualitative basis. For example, different medical imaging modalities can be jointly visualized for a given individual to identify common spatial patterns of signal changes. Data integration can be subsequently performed by jointly exploring unimodal features and unimodal analysis results. To this end, we may stratify the cohort of a clinical study based on some biomarkers extracted from different medical imaging modalities exceeding predefined thresholds. Finally, multivariate statistical and machine learning techniques can be applied for data-driven analysis of the joint relationship between information encoded in different modalities. Such approaches attempt to maximize the advantages of combining cross-modality information, dimensions, and resolution of the multimodal signal. The ultimate goal of such analysis methods is to identify the “mechanisms” underlying the generation of the observed medical data, to provide a joint representation of the common variation of heterogeneous data types.

The literature on multimodal analysis approaches is extensive, depending on the kind of applications and related data types. In this chapter we focus on general data integration methods, which can be classically related to the fields of *multivariate statistical analysis* and *latent variable modeling*. The importance of these approaches lies in the generality of their formulation, which makes them an ideal baseline for the analysis of heterogeneous data types. Furthermore, this chapter illustrates current extensions of these basic approaches to deep probabilistic models, which allow great modeling flexibility for current state-of-the-art applications.

In Subheading 1.2 we provide an overview of typical multimodal analyses in neuroimaging applications, while in Subheading 2 we introduce the statistical foundations of multivariate latent variable modeling, with emphasis on the standard approaches of partial least squares (PLS) and canonical correlation analysis (CCA). In Subheading 3, these classical methods are reformulated under the Bayesian lens, to define linear counterparts of latent variable models (Subheading 3.2) and their extension to multi-channel and deep multivariate analysis (Subheadings 3.3 and 3.4). In Subheading 4 we finally address the problem of group-wise regularization to improve the interpretability of multivariate association models, with specific focus in imaging-genetics applications.

Box 1: Online Tutorial

The material covered in this chapter is available at the following online tutorial:

<https://bit.ly/3y4RaIO>

1.2 Motivation from Neuroimaging Applications

Multimodal analysis methods have been explored for their potential in automatic patient diagnosis and stratification, as well as for their ability to identify interpretable data patterns characterizing clinical conditions. In this section, we summarize state-of-the-art contributions to the field, along with the remaining challenges to improve our understanding and applications to complex brain disorders.

- *Structural-structural combination.* Methods combining sMRI and dMRI imaging modalities are predominant in the field. Such combined analysis has been proposed, for example, for the detection of brain lesions (e.g., strokes [6, 7]) and to study and improve the management of patients with brain disorders [8].
- *Functional-functional combination.* Due to the complementary nature of EEG and fMRI, research in brain connectivity analysis has focused in the fusion of these modalities, to optimally integrate the high temporal resolution of EEG with the high spatial resolution of the fMRI signal. As a result, EEG-fMRI can provide simultaneous cortical and subcortical recording of brain activity with high spatiotemporal resolution. For example, this combination is increasingly used to provide clinical support for the diagnosis and treatment of epilepsy, to accurately localize seizure onset areas, as well as to map the surrounding functional cortex in order to avoid disability [9–11].
- *Structural-functional combination.* The combined analysis of sMRI, dMRI, and fMRI has been frequently proposed in neuropsychiatric research due to the high clinical availability of these imaging modalities and due to their potential to link brain function, structure, and connectivity. A typical application is in the study of autism spectrum disorder and attention-deficit hyperactivity disorder (ADHD). The combined analysis of such modalities has been proposed, for example, for the identification of altered white matter connectivity patterns in children with ADHD [12], highlighting association patterns between regional brain structural and functional abnormalities [13].
- *Imaging-genetics.* The combination of imaging and genetics data has been increasingly studied to identify genetic risk factors (genetic variations) associated with functional or structural abnormalities (quantitative traits, QTs) in complex brain disorders [3]. Such multimodal analyses are key to identify the

underlying mechanisms (from genotype to phenotype) leading to neurodegenerative diseases, such as Alzheimer's disease [14] or Parkinson's disease [15]. This analysis paradigm paves the way to novel data integration scenarios, including imaging and transcriptomics, or multi-omic data [16].

Overall, multimodal data integration in the study of brain disorders has shown promising results and is an actively evolving field. The potential of neuroimaging information is continuously improving, with increasing resolution and improved image contrast. Moreover, multiple imaging modalities are increasingly available in large collections of multimodal brain data, allowing for the application of complex modeling approaches on representative cohorts.

2 Methodological Background

2.1 From Multivariate Regression to Latent Variable Models

The use of multivariate analysis methods for biomedical data analysis is widespread, for example, in neuroscience [17], genetics [18], and imaging-genetics studies [19, 20]. These approaches come with the potential of explicitly highlighting the underlying relationship between data modalities, by identifying sets of relevant features that are jointly associated to explain the observed data.

In what follows, we represent the multimodal information available for a given subject k as a collection of arrays \mathbf{x}_i^k , $i=1, \dots, M$, where M is the number of available modalities. Each array has dimension $\dim(\mathbf{x}_i^k) = D_i$. A multimodal data matrix for N individuals is therefore represented by the collection of matrices \mathbf{X}_i , with $\dim(\mathbf{X}_i) = N \times D_i$. For sake of simplicity, we assume that $\mathbf{x}_i^k \in \mathbb{R}^{D_i}$.

A first assumption that can be made for defining a multivariate analysis method is that a target modality, say \mathbf{X}_j , is generated by the combination of a set of given modalities $\{\mathbf{X}_i\}_{i \neq j}$. A typical example of this application concerns the prediction of certain clinical variables from the combination of imaging features. In this case, the underlying forward *generative model* for an observation \mathbf{x}_j^k can be expressed as:

$$\mathbf{x}_j^k = g(\{\mathbf{x}_i^k\}_{i \neq j}) + \boldsymbol{\epsilon}_j^k, \quad (1)$$

where we assume that there exists an ideal mapping $g(\cdot)$ that transforms the ensemble of observed modalities for the individual k , to generate the target one \mathbf{x}_j^k . Note that we generally assume that the observations are corrupted by a certain noise $\boldsymbol{\epsilon}_j^k$, whose nature depends on the data type. The standard choice for the noise is Gaussian, $\boldsymbol{\epsilon}_j^k \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{Id})$.

Within this setting, a multimodal model is represented by a function $f(\{\mathbf{X}_i\}_{i=1}^M, \boldsymbol{\theta})$, with parameters $\boldsymbol{\theta}$, taking as input the ensemble of modalities across subjects. The model f is optimized

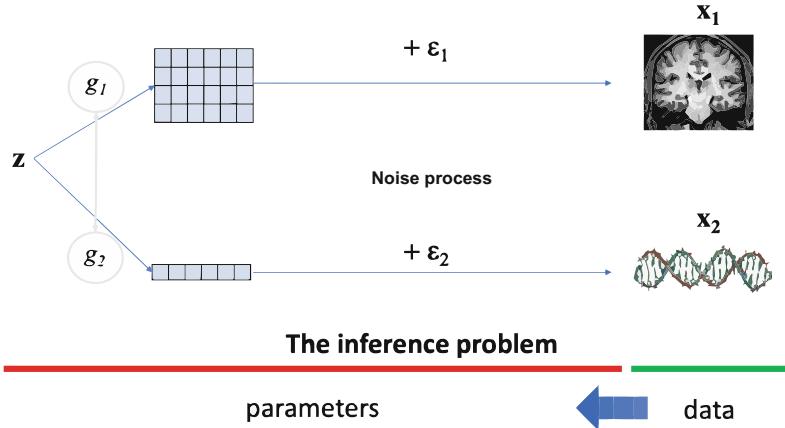


Fig. 1 Illustration of a generative process for the modeling of imaging and genetics data

with respect to θ to solve a specific task. In our case, the set of input modalities can be used to predict a target modality j , in this case we have $f : \bigotimes_{i \neq j} \mathbb{R}^{D_i} \mapsto \mathbb{R}^{D_j}$.

In its basic form, this kind of formulation includes standard multivariate linear regression, where the relationship between two modalities X_1 and X_2 is modeled through a set of linear parameters $\theta = W \in \mathbb{R}^{D_2 \times D_1}$ and $f(X_2) = X_2 \cdot W$. Under the Gaussian noise assumption, the typical optimization task is formulated as the least squares problem:

$$W^* = \underset{W}{\operatorname{argmin}} \|X_1 - X_2 \cdot W\|^2. \quad (2)$$

When modeling jointly multiple modalities, the forward generative model of Eq. 1 may be suboptimal, as it implies the explicit dependence of the target modality upon the other ones. This assumption may be too restrictive, as often an explicit assumption of dependency cannot be made, and we are rather interested in modeling the *joint variation* between data modalities. This is the rationale of *latent variable models*.

In the latent variable setting, we assume that the multiple modalities are jointly dependent from a common latent representation z (Fig. 1) belonging to an ideal low-dimensional space of dimension $D \leq \min\{\dim(D_i), i=1, \dots, M\}$.¹ In this case, Eq. 1 can be extended to the generative process:

$$\mathbf{x}_i^k = g_i(z_k) + \epsilon_i^k, \quad i=1, \dots, M. \quad (3)$$

¹ Note that we could also consider *overcomplete* basis for the latent space such that $D > \min\{\dim(D_i), i=1, \dots, M\}$. This choice may be motivated by the need of accounting for modalities with particularly low dimension. The study of overcomplete latent data representations is focus of active research [21–23].

Equation 3 is the forward process governing the data generation. The goal of latent variable modeling is to make inference on the latent space and on the generative process from the observed data modalities, based on specific assumptions on the transformations from the latent to the data space, and on the kind of noise process affecting the observations (Box 2). In particular, the inference problem can be tackled by estimating inverse mappings, $f_j(\mathbf{x}_j^k)$, from the data space of the observed modalities to the latent space.

Based on this framework, in the following sections, we illustrate the standard approaches for solving the inference problem of Eq. 1.

Box 2: Online Tutorial—Generative Models

The forward model of Eq. 3 for multimodal data generation can be easily coded in Python to generate a synthetic multimodal dataset:

```
# N subjects
n = 500
# here we define 2 Gaussian latents variables
# z = (l_1, l_2)
l1 = np.random.normal(size=n)
l2 = np.random.normal(size=n)

latents = np.array([l1, l2]).T

# We define two random transformations from the latent
# space to the 5D space of X1 and X2 respectively
transform_x = \
    np.random.randint(-8,8, size = 10).reshape([2,5])
transform_y = \
    np.random.randint(-8,8, size = 10).reshape([2,5])

# We compute data X = z w_x, and Y = z w_y
X1 = latents.dot(transform_x)
X2 = latents.dot(transform_y)

# We add some random Gaussian noise
X1 = X1 + 2*np.random.normal(size = n*5).reshape((n, 5))
X2 = X2 + 2*np.random.normal(size = n*5).reshape((n, 5))
```

2.2 Classical Latent Variable Models: PLS and CCA

Classical latent variable models extend the standard linear regression to analyze the joint variability of different modalities. Typical formulation of latent variable models include *partial least squares* (PLS) and *canonical correlation analysis* (CCA) [24], which have successfully been applied in biomedical research [25], along with multimodal [26, 27] and nonlinear [28, 29] variants.

Box 3: Online Tutorial—PLS and CCA with sklearn

```

from sklearn.cross_decomposition import PLSCanonical, CCA

#####
# We fit PLS and CCA as provided by scikit-learn

#Defining PLS object, no scaling of input X1 and X2
plsca = PLSCanonical(n_components=2, scale = False)
cca = CCA(n_components=2, scale = False)

#Fitting on train data
plsca.fit(X1, X2)
cca.fit(X1, X2)

#We project the training data in the latent dimension
X1_pls_r, X2_pls_r = \
    plsca.transform(X1, X2)
X1_cca_r, X2_cca_r = \
    cca.transform(X1, X2)

```

The basic principle of these multivariate analysis techniques relies on the identification of *linear transformations* of modalities X_i and X_j into a lower dimensional subspace of dimension $D \leq \min\{\dim(D_i), \dim(D_j)\}$, where the projected data exhibits the desired statistical properties of similarity. For example, PLS aims at maximizing the covariance between these combinations (or projections on the modes' directions), while CCA maximizes their statistical correlation (Box 3). For simplicity, in what follows we focus on the joint analysis of two modalities X_1 and X_2 , and the multimodal model can be written as

$$f(X_1, X_2, \theta) = [f_1(X_1, u_1), f_2(X_2, u_2)] \quad (4)$$

$$= [z_1, z_2], \quad (5)$$

where $\theta = \{u_1, u_2\}$ are linear projection operators for the modalities, $u_i \in \mathbb{R}^{D_i}$, while $z_i = X_i \cdot u_i \in \mathbb{R}^N$ are the latent projections for each modality $i = 1, 2$. The optimization problem can thus be formulated as:

$$u_1^*, u_2^* = \underset{\theta}{\operatorname{argmax}} \quad Sim(z_1, z_2) \quad (6)$$

$$= \underset{u_1, u_2}{\operatorname{argmax}} \quad Sim(X_1 \cdot u_1, X_2 \cdot u_2), \quad (7)$$

where Sim is a suitable measure of statistical similarity, depending on the envisaged methods (e.g., variance for PLS, or correlation for CCA) (Fig. 2).

Latent variable modeling

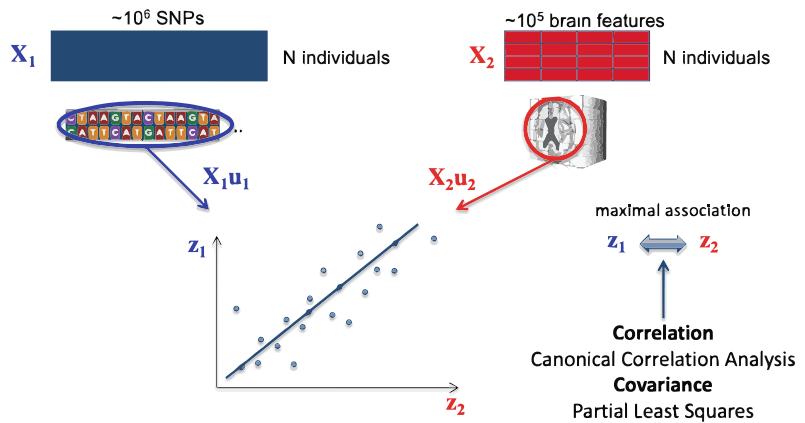


Fig. 2 Illustration of latent variable modeling for an idealized application to the modeling of genetics and imaging data

2.3 Latent Variable Models Through Eigen-Decomposition

2.3.1 Partial Least Squares

For PLS, the problem of Eq. 6 requires the estimation of projections \mathbf{u}_1 and \mathbf{u}_2 maximizing the *covariance* between the latent representation of the two modalities X_1 and X_2 :

$$\mathbf{u}_1^*, \mathbf{u}_2^* = \underset{\mathbf{u}_1, \mathbf{u}_2}{\operatorname{argmax}} \operatorname{Cov}(\mathbf{X}_1 \cdot \mathbf{u}_1, \mathbf{X}_2 \cdot \mathbf{u}_2), \quad (8)$$

where

$$\operatorname{Cov}(\mathbf{X}_1 \cdot \mathbf{u}_1, \mathbf{X}_2 \cdot \mathbf{u}_2) = \frac{\mathbf{u}_1^T \mathbf{S} \mathbf{u}_2}{\sqrt{\mathbf{u}_1^T \mathbf{u}_1} \sqrt{\mathbf{u}_2^T \mathbf{u}_2}}, \quad (9)$$

and $\mathbf{S} = \mathbf{X}_1^T \mathbf{X}_2$ is the sample covariance between modalities.

Without loss of generality, the maximization of Eq. 9 can be considered under the orthogonality constraint $\sqrt{\mathbf{u}_1^T \mathbf{u}_1} = \sqrt{\mathbf{u}_2^T \mathbf{u}_2} = 1$. This constrained optimization problem can be expressed in the Lagrangian form:

$$\mathcal{L}(\mathbf{u}_1, \mathbf{u}_2, \lambda_x, \lambda_y) = \mathbf{u}_1^T \mathbf{S} \mathbf{u}_2 - \lambda_x (\mathbf{u}_1^T \mathbf{u}_1 - 1) - \lambda_y (\mathbf{u}_2^T \mathbf{u}_2 - 1), \quad (10)$$

whose solution can be written as:

$$\begin{bmatrix} \mathbf{0} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}. \quad (11)$$

Equation 11 corresponds to the *primal* formulation of PLS and shows that the PLS projections maximizing the latent covariance are the left and right eigen-vectors of the sample covariance matrix across modalities. This solution is known as PLS-SVD and has been widely adopted in the field of neuroimaging [30, 31], for the study of common patterns of variability between multimodal imaging data, such as PET and fMRI.

It is worth to notice that classical *principal component analysis* (PCA) is a special case of PLS when $\mathbf{X}_1 = \mathbf{X}_2$. In this case the latent projections maximize the data variance and correspond to the eigen-modes of the sample covariance matrix $\mathbf{S} = \mathbf{X}_1^T \mathbf{X}_1$.

2.3.2 Canonical Correlation Analysis

In canonical correlation analysis (CCA), the problem of Eq. 6 is formulated by optimizing linear transformations such that $\mathbf{X}_1 \mathbf{u}_1$ and $\mathbf{X}_2 \mathbf{u}_2$ are maximally correlated:

$$\mathbf{u}_1^*, \mathbf{u}_2^* = \underset{\mathbf{u}_1, \mathbf{u}_2}{\operatorname{argmax}} \operatorname{Corr}(\mathbf{X}_1 \mathbf{u}_1, \mathbf{X}_2 \mathbf{u}_2), \quad (12)$$

where

$$\operatorname{Corr}(\mathbf{X}_1 \mathbf{u}_1, \mathbf{X}_2 \mathbf{u}_2) = \frac{\mathbf{u}_1^T \mathbf{S} \mathbf{u}_2}{\sqrt{\mathbf{u}_1^T \mathbf{S}_1 \mathbf{u}_1} \sqrt{\mathbf{u}_2^T \mathbf{S}_2 \mathbf{u}_2}}. \quad (13)$$

where $\mathbf{S}_1 = \mathbf{X}_1^T \mathbf{X}_1$ and $\mathbf{S}_2 = \mathbf{X}_2^T \mathbf{X}_2$ are the sample covariances of modality 1 and 2, respectively.

Proceeding in a similar way as for the derivation of PLS, it can be shown that CCA is associated to the generalized eigen-decomposition problem [32]:

$$\begin{bmatrix} \mathbf{0} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}, \quad (14)$$

It is common practice to reformulate the CCA problem of Eq. 14 with a *regularized* version aimed to avoid numerical instabilities due to the estimation of the sample covariances \mathbf{S}_1 and \mathbf{S}_2 :

$$\begin{bmatrix} \mathbf{0} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{S}_1 + \delta \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 + \delta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}. \quad (15)$$

In this latter formulation, the right hand side of Eq. 14 is regularized by introducing a constant diagonal term δ , proportional to the regularization strength (with $\delta=0$ we obtain Eq. 14). Interestingly, for large value of δ , the diagonal term dominates the sample covariance matrices of the right-hand side, and we retrieve the standard eigen-value problem of Eq. 11. This shows that PLS can be interpreted as an infinitely regularized formulation of CCA.

2.4 Kernel Methods for Latent Variable Models

In order to capture nonlinear relationships, we may wish to project our input features into a high-dimensional space prior to performing CCA (or PLS):

$$\phi : \mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^N) \mapsto [\phi(\mathbf{x}^1), \dots, \phi(\mathbf{x}^N)] \quad (16)$$

where ϕ is a nonlinear feature map. As derived by Bach et al. [33], the data matrices \mathbf{X}_1 and \mathbf{X}_2 can be replaced by the Gram matrices \mathbf{K}_1 and \mathbf{K}_2 such that we can achieve a nonlinear feature mapping via the kernel trick [34]:

$$\mathbf{K}_1(\mathbf{x}_1^i, \mathbf{x}_1^j) = \langle \phi(\mathbf{x}_1^i), \phi(\mathbf{x}_1^j) \rangle \text{ and } \mathbf{K}_2(\mathbf{x}_2^i, \mathbf{x}_2^j) = \langle \phi(\mathbf{x}_2^i), \phi(\mathbf{x}_2^j) \rangle \quad (17)$$

where $\mathbf{K}_1 = [\mathbf{K}_1(\mathbf{x}_1^i, \mathbf{x}_1^j)]_{N \times N}$ and $\mathbf{K}_2 = [\mathbf{K}_2(\mathbf{x}_2^i, \mathbf{x}_2^j)]_{N \times N}$. In this case, kernel CCA canonical directions correspond to the solutions of the updated generalized eigen-value problem:

$$\begin{bmatrix} \mathbf{0} & \mathbf{K}_1 \mathbf{K}_2 \\ \mathbf{K}_2 \mathbf{K}_1 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{K}_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_2^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}. \quad (18)$$

Similarly to the primal formulation of CCA, we can apply an ℓ_2 -norm regularization penalty on the weights α_1 and α_2 of Eq. 18, giving rise to regularized kernel CCA:

$$\begin{bmatrix} \mathbf{0} & \mathbf{K}_1 \mathbf{K}_2 \\ \mathbf{K}_2 \mathbf{K}_1 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{K}_1^2 + \delta I & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_2^2 + \delta I \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}, \quad (19)$$

2.5 Optimization of Latent Variable Models

The *nonlinear iterative partial least squares* (NIPALS) is a classical scheme proposed by H. Wold [35] for the optimization of latent variable models through the iterative computation of PLS and CCA projections. Within this method, the projections associated with the modalities \mathbf{X}_1 and \mathbf{X}_2 are obtained through the iterative solution of simple least squares problems.

The principle of NIPALS is to identify projection vectors $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}$ and corresponding latent representations \mathbf{z}_1 and \mathbf{z}_2 to minimize the functionals

$$\mathcal{L}_i = \|X_i - \mathbf{z}_i \mathbf{u}_i^T\|^2, \quad (20)$$

subject to the constraint of maximal similarity between representations \mathbf{z}_1 and \mathbf{z}_2 (Fig. 3).

Following [37], the NIPALS method is optimized as follows (Algorithm 1). The latent projection for modality 1 is first initialized as $\mathbf{z}_1^{(0)}$ from randomly chosen columns of the data matrix \mathbf{X}_1 . Subsequently, the linear regression function

$$\mathcal{L}_2^{(0)} = \|X_2 - \mathbf{z}_1^{(0)} \mathbf{u}_2^T\|^2$$

is optimized with respect to \mathbf{u}_2 , to obtain the projection $\mathbf{u}_2^{(0)}$. After unit scaling of the projection coefficients, the new latent representation is computed for modality 2 as $\mathbf{z}_2^{(0)} = \mathbf{X}_2 \cdot \mathbf{u}_2^{(0)}$. At this point, the latent projection is used for a new optimization step of the linear regression problem

$$\mathcal{L}_1^{(0)} = \|X_1 - \mathbf{z}_2^{(0)} \mathbf{u}_1^T\|^2,$$

this time with respect to \mathbf{u}_1 , to obtain the projection parameters $\mathbf{u}_1^{(0)}$ relative to modality 1. After unit scaling of the coefficients, the new latent representations is computed for modality 1 as $\mathbf{z}_1^{(1)} = \mathbf{X}_1 \cdot \mathbf{u}_1^{(0)}$. The whole procedure is then iterated.

Non-linear iterative partial least squares - NIPALS

[scikit-learn/sklearn/cross_decomposition](#)

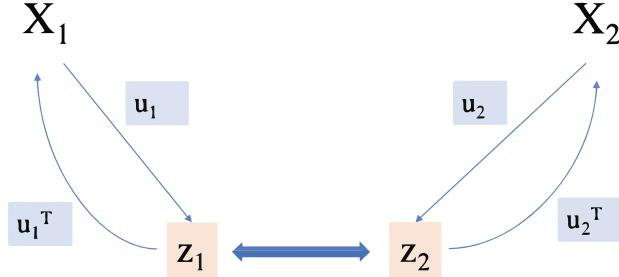


Fig. 3 Schematic of NIPALS algorithm (Algorithm 1). This implementation can be found in standard machine learning packages such as scikit-learn [36]

It can be shown that the NIPALS method of Algorithm 1 converges to a stable solution for projections and latent parameters and the resulting projection vectors correspond to the first left and right eigen-modes associated to the covariance matrix $S = \mathbf{X}_1^T \cdot \mathbf{X}_2$.

Algorithm 1 NIPALS iterative computation for PLS components [37]

Initialize $\mathbf{z}_1^{(0)}$, $i = 0$.

Until not converged do:

1. Estimate the projection $\mathbf{u}_2^{(i)}$ by minimizing $\mathcal{L}_2^{(i)} = \|\mathbf{X}_2 - \mathbf{z}_1^{(i)} \mathbf{u}_2^T\|^2$:

$$\mathbf{u}_2^{(i)} = \mathbf{X}_2^T \mathbf{z}_1^{(i)} \left(\mathbf{z}_1^{(i)T} \mathbf{z}_1^{(i)} \right)^{-1}$$
2. Normalize $\mathbf{u}_2^{(i)} \leftarrow \frac{\mathbf{u}_2^{(i)}}{\|\mathbf{u}_2^{(i)}\|}$.
3. Estimate the latent representation for modality 2:

$$\mathbf{z}_2^{(i)} = \mathbf{X}_2 \cdot \mathbf{u}_2^{(i)}$$
4. Estimate the projection $\mathbf{u}_1^{(i)}$ by minimizing $\mathcal{L}_1^{(i)} = \|\mathbf{X}_1 - \mathbf{z}_2^{(i)} \mathbf{u}_1^T\|^2$:

$$\mathbf{u}_1^{(i)} = \mathbf{X}_1^T \mathbf{z}_2^{(i)} \left(\mathbf{z}_2^{(i)T} \mathbf{z}_2^{(i)} \right)^{-1}$$
5. Normalize $\mathbf{u}_1^{(i)} \leftarrow \frac{\mathbf{u}_1^{(i)}}{\|\mathbf{u}_1^{(i)}\|}$.
6. Update the latent representation for modality 1:

$$\mathbf{z}_1^{(i+1)} = \mathbf{X}_1 \cdot \mathbf{u}_1^{(i)}$$
.

After the first eigen-modes are computed through Algorithm 1, the higher-order components can be subsequently computed by *deflating* the data matrices \mathbf{X}_1 and \mathbf{X}_2 . This can be done by regressing out the current projections in the latent space:

$$\mathbf{X}_i \leftarrow \mathbf{X}_i - \mathbf{z}_i \frac{\mathbf{z}_i^T \mathbf{X}_i}{\mathbf{z}_i^T \mathbf{z}_i} \quad (21)$$

NIPALS can be seamlessly used to optimize the CCA problem. Indeed, it can be shown that the CCA projections and latent representations can be obtained by estimating the linear projections \mathbf{u}_2 and \mathbf{u}_1 in **steps 1** and **4** of Algorithm 1 via the linear regression problems

$$\mathcal{L}_2^{(i)} = \|X_2 \mathbf{u}_2 - \mathbf{z}_1^{(i)}\|^2 \quad (\text{step 1 for CCA}),$$

and

$$\mathcal{L}_1^{(i)} = \|X_1 \mathbf{u}_1 - \mathbf{z}_2^{(i)}\|^2 \quad (\text{step 4 for CCA}).$$

Box 4: Online Tutorial—NIPALS Implementation

The online tutorial provides an implementation of the NIPALS algorithm for both CCA and PLS, corresponding to Algorithm 1. It can be verified that the numerical solution is equivalent to the one provided by sklearn and to the one obtained through the solution of the eigen-value problem.

3 Bayesian Frameworks for Latent Variable Models

Bayesian formulations for latent variable models have been developed in the past, including for PLS [38] and CCA [39]. The advantage of employing a Bayesian framework to solve the original inference problem is that it provides a natural setting to quantify the parameters' variability in an interpretable manner, coming with their estimated distribution. In addition, these methods are particularly attractive for their ability of integrating prior knowledge on the model's parameters.

3.1 Multi-view PPCA

Recently, the seminal work of Tipping and Bishop on probabilistic PCA (PPCA) [40] has been extended to allow the joint integration of multimodal data [41] (*multi-view PPCA*), under the assumption of a common latent space able to explain and generate all modalities.

Recalling the notation of Subheading 2.1, let $\mathbf{x} = \{\mathbf{x}_i^k\}_{i=1}^M$ be an observation of M modalities for subject k , where each \mathbf{x}_i^k is a vector of dimension D_i . We denote by \mathbf{z}^k the D -dimensional latent variable commonly shared by each \mathbf{x}_i^k . In this context, the forward process underlying the data generation of Eq. 1 is linear, and for each subject k and modality i , we write (see Fig. 4a):

$$\mathbf{x}_i^k = W_i(\mathbf{z}^k) + \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i, \quad (22)$$

$$i = 1, \dots, M; \quad k = 1, \dots, N; \quad \dim(\mathbf{z}^k) < \min(D_i), \quad (23)$$

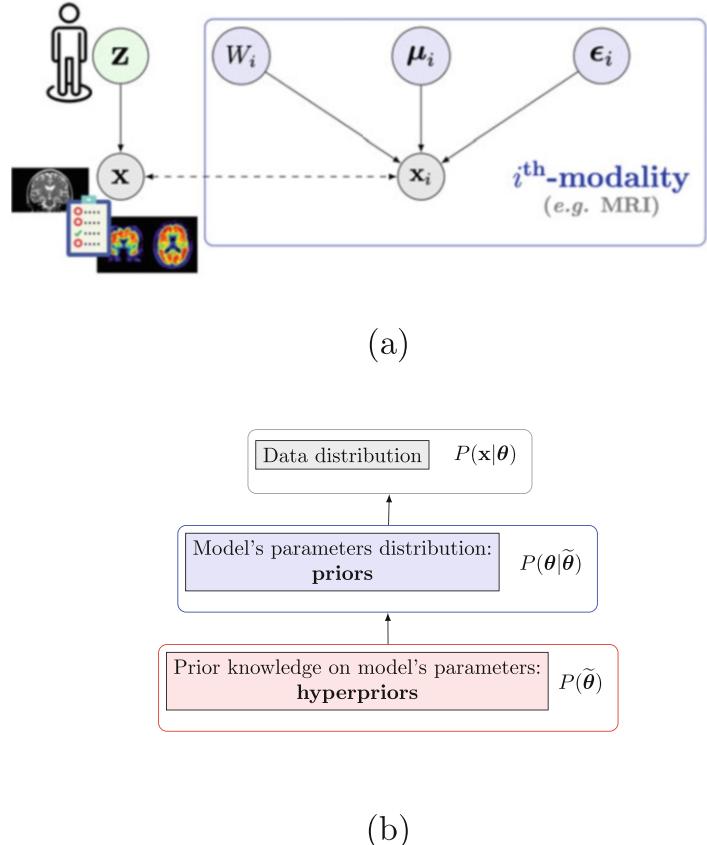


Fig. 4 (a) Graphical model of multi-view PPCA. The green node represents the latent variable able to jointly describe all observed data explaining the patient status. Gray nodes denote original multimodal data, and blue nodes the view-specific parameters. (b) Hierarchical structure of multi-view PPCA: prior knowledge on model's parameters can be integrated in a natural way when the model is embedded in a Bayesian framework

where W_i represents the linear mapping from the i th-modality to the latent space, while $\boldsymbol{\mu}_i$ and $\boldsymbol{\epsilon}_i$ denote the common intercept and error for modality i . Note that the modality index i does not appear in the latent variable \mathbf{z}^k , allowing a compact formulation of the generative model of the whole dataset (i.e., including all modalities) by simple concatenation:

$$\mathbf{x}^k := \begin{bmatrix} \mathbf{x}_1^k \\ \vdots \\ \mathbf{x}_M^k \end{bmatrix} = \begin{bmatrix} W_1 \\ \vdots \\ W_M \end{bmatrix} \mathbf{z}^k + \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_M \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_M \end{bmatrix} = : W\mathbf{z}^k + \boldsymbol{\mu} + \boldsymbol{\epsilon}. \quad (24)$$

Further hypotheses are needed to define the probability distributions of each element appearing in Eq. 22, such as $\mathbf{z}^k \sim p(\mathbf{z}^k)$, the

standard Gaussian prior distribution for the latent variables, and $\boldsymbol{\epsilon}_i \sim p(\boldsymbol{\epsilon}_i)$, a centered Gaussian distribution. From these assumptions, one can finally derive the likelihood of the data given latent variables and model parameters, $p(\mathbf{x}_i^k | \mathbf{z}^k, \boldsymbol{\theta}_i)$, $\boldsymbol{\theta}_i = \{W_i, \boldsymbol{\mu}_i, \boldsymbol{\epsilon}_i\}$ and, by using Bayes theorem, also the posterior distribution of the latent variables, $p(\mathbf{z}^k | \mathbf{x}_i^k)$.

Box 5: Online Tutorial—Multi-view PPCA

```
from Model.mvPPCA import MVPPCA

### Data in mv-PPCA is specified by:
# 1 - number of views, views' dimensions
# and latent dimension
n_views = 2 # X1 and X2
n_components = n_components
dim_views = [X.shape[1], Y.shape[1]]
# 2 - a dataframe containing all views
data = pd.DataFrame(np.hstack((X, Y)))

### Here we create an instance of the model
#and a dataframe to store results during training
n_iterations=200
results = pd.DataFrame()
# Multi-views PCA
mvPPCA = MVPPCA(data=data, norm=False,
                  dim_views=dim_views,
                  n_components=n_components,
                  n_iterations=n_iterations)
#####
## Model Fitting ##
#####
results = results.append(mvPPCA.fit(), ignore_index=True)
# Optimized parameters can be recovered as follows:
muk, Wk, Sigma2k = mvPPCA.local_params
```

3.1.1 Optimization

In order to solve the inference problem and estimate the model's parameters in $\boldsymbol{\theta}$, the classical expectation-maximization (EM) scheme can be deployed. EM optimization consists in an iterative process where each iteration is composed of two steps:

- Expectation step (E): Given the parameters previously optimized, the expectation of the log-likelihood of the joint distribution of \mathbf{x}_i and \mathbf{z}^k with respect to the posterior distribution of the latent variables is evaluated.
- Maximization step (M): The functional of the E step is maximized with respect to the model's parameters.

It is worth noticing that prior knowledge on the model's parameters distribution can be easily integrated in this Bayesian framework (Fig. 4b), with minimal modification of the optimization scheme, consisting in a penalization of the functional to be maximized in the M-step forcing the optimized parameters to remain close to their priors. In this case we talk about maximum a posteriori (MAP) optimization.

3.2 Bayesian Latent Variable Models via Autoencoding

Autoencoders and variational autoencoders have become very popular approaches for the estimation of latent representation of complex data, which allow powerful extensions of the Bayesian models presented in Subheading 3.1 to account for nonlinear and deep data representations.

Autoencoders (AEs) extend classical latent variable models to account for complex, potentially highly nonlinear, projections from the data space to the latent space (encoding), along with reconstruction functions (decoding) mapping the latent representation back to the data space. Since typical encoding (f_e) and decoding (f_d) functions of AEs are parameterized by feedforward neural networks, inference can be efficiently performed by means of stochastic gradient descent through backpropagation. In this sense, AEs can be seen as a powerful extension of classical PCA, where encoding into the latent representations and decoding are jointly optimized to minimize the reconstruction error of the data:

$$\mathcal{L} = \| \mathbf{X} - f_d(f_e(\mathbf{X})) \|_2^2 \quad (25)$$

The *variational autoencoder* (VAE) [42, 43] introduces a Bayesian formulation of AEs, akin to PPCA, where the latent variables are inferred by estimating the associated posterior distributions. In this case, the optimization problem can be efficiently performed by *stochastic variational inference* [44], where the posterior moments of the variational posterior of the latent distribution are parameterized by neural networks.

In the same way PLS and CCA extend PCA for multimodal analysis, research has been devoted to define equivalent extensions for the VAEs to identify common latent representations of multiple data modalities, such as the multi-channel VAE [23], or deep CCA [29]. These approaches are based on a similar formulation, which is provided in the following section.

3.3 Multi-channel Variational Autoencoder

The multi-channel variational autoencoder (mcVAE) assumes the following generative process for the observation set:

$$\begin{aligned} \mathbf{z}^k &\sim p(\mathbf{z}^k) \\ \mathbf{x}_i^k &\sim p(\mathbf{x}_i^k | \mathbf{z}^k, \theta_i) \quad i = 1, \dots, M, \end{aligned} \tag{26}$$

where $p(\mathbf{z}^k)$ is a prior distribution for the latent variable. In this case, $p(\mathbf{x}_i^k | \mathbf{z}, \theta_i)$ is the likelihood of the observed modality i for subject k , conditioned on the latent variable and on the generative parameters θ_i parameterizing the decoding from the latent space to the data space of modality i .

Solving this inference problem requires the estimation of the posterior for the latent distribution $p(\mathbf{z} | \mathbf{X}_1, \dots, \mathbf{X}_M)$, which is generally an intractable problem. Following the VAE scheme, *variational inference* can be applied to compute an approximate posterior [45].

3.3.1 Optimization

The inference problem of mcVAE is solved by identifying variational posterior distributions specific to each data modality $q(\mathbf{z}^k | \mathbf{x}_i^k, \varphi_i)$, by conditioning them on the observed modality \mathbf{x}_i and on the corresponding variational parameters φ_i parameterizing the encoding of the observed modality to the latent space.

In this way, since each modality provides a different approximation, a similarity constraint is imposed in the latent space to enforce each modality-specific distribution $q(\mathbf{z}^k | \mathbf{x}_i^k, \varphi_i)$ to be as close as possible to the common target posterior distribution. The measure of “proximity” between distributions is the Kullback-Leibler (KL) divergence. This constraint defines the following functional:

$$\underset{q}{\operatorname{argmin}} \sum_i D_{\text{KL}}[q(\mathbf{z}^k | \mathbf{x}_i^k, \varphi_i) \| p(\mathbf{z} | \mathbf{x}_1^k, \dots, \mathbf{x}_M^k)] \tag{27}$$

where the approximate posteriors $q(\mathbf{z} | \mathbf{x}_i, \varphi_i)$ represent the view on the latent space that can be inferred from the modality \mathbf{x}_i . In [23] it was shown that the optimization of Eq. 27 is equivalent to the optimization of the following evidence lower bound (ELBO):

$$\mathcal{L} = D - R \tag{28}$$

where $R = \sum_i \text{KL}[q(\mathbf{z}^k | \mathbf{x}_i^k, \varphi_i) \| p(\mathbf{z})]$, and $D = \sum_i L_i$, with

$$L_i = \mathbb{E}_{q(\mathbf{z}^k | \mathbf{x}_i^k, \varphi_i)} \sum_{j=1}^M \ln p(\mathbf{x}_j | \mathbf{z}, \theta_j)$$

is the expected log-likelihood of each data channel \mathbf{x}_j quantifying the reconstruction obtained by decoding from the latent representation of the remaining channels \mathbf{x}_i . Therefore, optimizing the term D in Eq. 28 with respect to encoding and decoding parameters $\{\theta_i, \varphi_i\}_{i=1}^M$ identifies the optimal representation of each modality in

the latent space which can, on average, jointly reconstruct all the other channels. This term thus enforces a coherent latent representation across different modalities and is balanced by the regularization term R , which constrains the latent representation of each modality to the common prior $p(\mathbf{z})$. As for standard VAEs, encoding and decoding functions can be arbitrarily chosen to parameterize respectively latent distributions and data likelihoods. Typical choices for such functions are neural networks, which can provide extremely flexible and powerful data representation (Box 6). For example, leveraging the modeling capabilities of deep convolutional networks, mcVAE has been used in a recent cardiovascular study for the prediction of cardiac MRI data from retinal fundus images [46].

Box 6 [Online Tutorial](#)—mcVAE with PyTorch

```

import torch
from mcvae.models import Mcvae
from mcvae.models.utils import DEVICE, load_or_fit

### Data in mcvae is specified by:
# 1 - a dictionary with the data characteristics
init_dict = {
    'n_channels': 2, # X1 and X2
    'lat_dim': n_components,
    'n_feats': tuple([X1.shape[1], X2.shape[1]]),
}
# 2 - a list with the different data channels
data = []
data.append(torch.FloatTensor(X1))
data.append(torch.FloatTensor(X2))

# Here we create an instance of the model
adam_lr = 1e-2
n_epochs = 4000
# Multi-Channel VAE
torch.manual_seed(24)
model = Mcvae(**init_dict)
model.to(DEVICE)
#####
## Model Fitting ##
#####
model.optimizer = torch.optim.Adam(model.parameters(), \
                                    lr=adam_lr)
load_or_fit(model=model, data=data, epochs=n_epochs, \
            ptfile='model.pt', force_fit=FORCE_REFIT)

```

3.4 Deep CCA

The mcVAE uses neural network layers to learn nonlinear representations of multimodal data. Similarly, Deep CCA [29] provides an alternative to kernel CCA to learn nonlinear mappings of multimodal information. Deep CCA computes representations by passing two views through functions f_1 and f_2 with parameters θ_1 and θ_2 , respectively, which can be learnt by multilayer neural networks. The parameters are optimized by maximizing the correlation between the learned representations $f_1(\mathbf{X}_1; \theta_1)$ and $f_2(\mathbf{X}_2; \theta_2)$:

$$(\boldsymbol{\theta}_{1opt}, \boldsymbol{\theta}_{2opt}) = \operatorname{argmax} \operatorname{Corr}(f_1(\mathbf{X}_1; \boldsymbol{\theta}_1), f_2(\mathbf{X}_2; \boldsymbol{\theta}_2)) (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \quad (29)$$

In its classical formulation, the correlation objective given in Eq. 29 is a function of the full training set, and as such, mini-batch optimization can lead to suboptimal results. Therefore, optimization of classical deep CCA must be performed with full-batch optimization, for example, through the L-BFGS (limited Broyden-Fletcher-Goldfarb-Shanno) scheme [47]. For this reason, with this vanilla implementation, deep CCA is not computationally viable for large datasets. Furthermore, this approach does not provide a model for generating samples from the latent space. To address these issues, Wang et al. [48] introduced deep variational CCA (VCCA) which extends the probabilistic CCA framework introduced in Subheading 3 to a nonlinear generative model. In a similar approach to VAEs and mcVAE, deep VCCA uses variational inference to approximate the posterior distribution and derives the following ELBO:

$$\mathcal{L} = -D_{\text{KL}}(q_{\phi}(\mathbf{z} | \mathbf{x}_1) \| p(\mathbf{z})) + \mathbb{E}_{q_{\phi}(\mathbf{z} | \mathbf{x}_1)} [\log p_{\theta_1}(\mathbf{x}_1 | \mathbf{z}) + \log p_{\theta_2}(\mathbf{x}_2 | \mathbf{z})] \quad (30)$$

where the approximate posterior, $q_{\phi}(\mathbf{z} | \mathbf{x}_1)$, and likelihood distributions, $p_{\theta_1}(\mathbf{x}_1 | \mathbf{z})$ and $p_{\theta_2}(\mathbf{x}_2 | \mathbf{z})$, are parameterized by neural networks with parameters ϕ , θ_1 , and θ_2 .

We note that, in contrast to mcVAE, deep VCCA is based on the estimation of a single latent posterior distribution. Therefore, the resulting representation is dependent on the reference modality from which the joint latent representation is encoded and may therefore bias the estimation of the latent representation. Finally Wang et al. [48] introduce a variant of deep VCCA, VCCA-private, which extracts the private, in addition to shared, latent information. Here, private latent variables hold view-specific information which is not shared across modalities.

4 Biologically Inspired Data Integration Strategies

Medical imaging and -omics data are characterized by nontrivial relationships across features, which represent specific mechanisms underlying the pathophysiological processes.

For example, the pattern of brain atrophy and functional impairment may involve brain regions according to the brain connectivity structure [49]. Similarly, biological processes such as gene expression are the result of the joint contribution of several SNPs acting according to *biological pathways*. According to these processes, it is possible to establish relationships between genetics features under the form of relation networks, represented by ontologies such as the KEGG pathways² and the Gene Ontology Consortium.³

When applying data-driven multivariate analysis methods to this kind of data, it is therefore relevant to promote interpretability and plausibility of the model, by enforcing the solution to follow the structural constraints underlying the data. This kind of model behavior can be achieved through *regularization* of the model parameters.

In particular, group-wise regularization [50] is an effective approach to enforce structural patterns during model optimization, where related features are jointly penalized with respect to a common parameter. For example, group-wise constraints may be introduced to account for biological pathways in models of gene association, or for known brain networks and regional interactions in neuroimaging studies. More specifically, we assume that the D_i features of a modality $\mathbf{x}_i = (x_{i1}, \dots, x_{iD_i})$ are grouped in subsets $\{S_l\}_{l=1}^L$, according to the indices $S_l = (s_1, \dots, s_{N_l})$. The regularization of the general multivariate model of Eq. 2 according to the group-wise constraint can be expressed as:

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmin}} \| \mathbf{X}_1 - \mathbf{X}_2 \cdot \mathbf{W} \|^2 + \lambda \sum_{l=1}^L \beta_l R(\mathbf{W}_l), \quad (31)$$

where $R(\mathbf{W}_l) = \sum_{j=1}^{D_l} \sqrt{\sum_{s \in S_l} \mathbf{W}[s, j]^2}$ is the penalization of the entries of \mathbf{W} associated with the features of \mathbf{X}_2 indexed by S_l . The total penalty is achieved by the sum across the D_l columns.

Group-wise regularization is particularly effective in the following situations:

- To compensate for large data dimensionality, by reducing the number of “free parameters” to be optimized by aggregating the available features [51].

² <https://www.genome.jp/kegg/pathway.html>.

³ <http://geneontology.org/>.

- To account for the small effect size of each independent features, to combine features in order to increase the detection power. For example, in genetic analysis, each SNP accounts for below 1% of the variance in brain imaging quantitative traits when considered individually [52, 53].
- To meaningfully integrate complementary information to introduce biologically inspired constraints into the model.

In the context of group-wise regularization in neural networks, several optimization/regularization strategies have been proposed to allow the identification of compressed representation of multimodal data in the bottleneck layers, such as by imposing sparsity of the model parameters or by introducing grouping constraints motivated by prior knowledge [54].

For instance, the Bayesian Genome-to-Phenome Sparse Regression (G2PSR) method proposed in [55] associates genomic data to phenotypic features, such as multimodal neuroimaging and clinical data, by constraining the transformation to optimize relevant group-wise SNPs-gene associations. The resulting architecture groups the input SNP layer into corresponding genes represented in the intermediate layer L of the network (Fig. 6). Sparsity at the gene level is introduced through variational dropout [56], to estimate the relevance of each gene (and related SNPs) in reconstructing the output phenotypic features.

In more detail, to incorporate biological constraints in G2PSR framework, a group-wise penalization is imposed with nonzero weights \mathbf{W}^g mapping the input SNPs to their common gene g . The idea is that during optimization the model is forced to jointly discard all the SNPs mapping to genes which are not relevant to the predictive task. Following [56], the variational approximation is

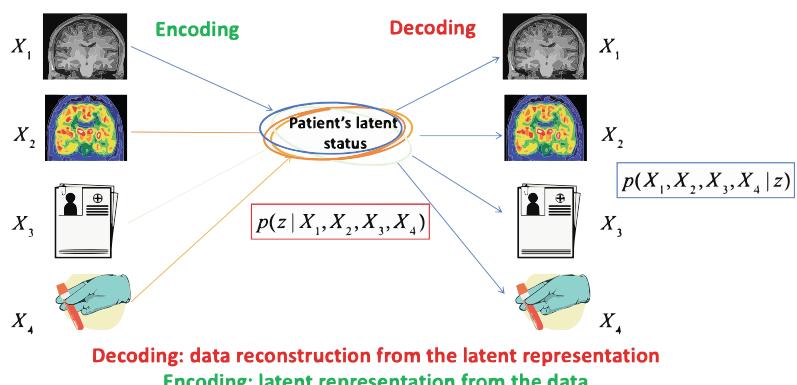


Fig. 5 The multi-channel VAE (mcVAE) for the joint modeling of multimodal medical imaging, clinical, and biological information. The mcVAE approximates the latent posterior $p(\mathbf{z}|\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4)$ to maximize the likelihood of the data reconstruction $p(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4|\mathbf{z})$ (plus a regularization term)

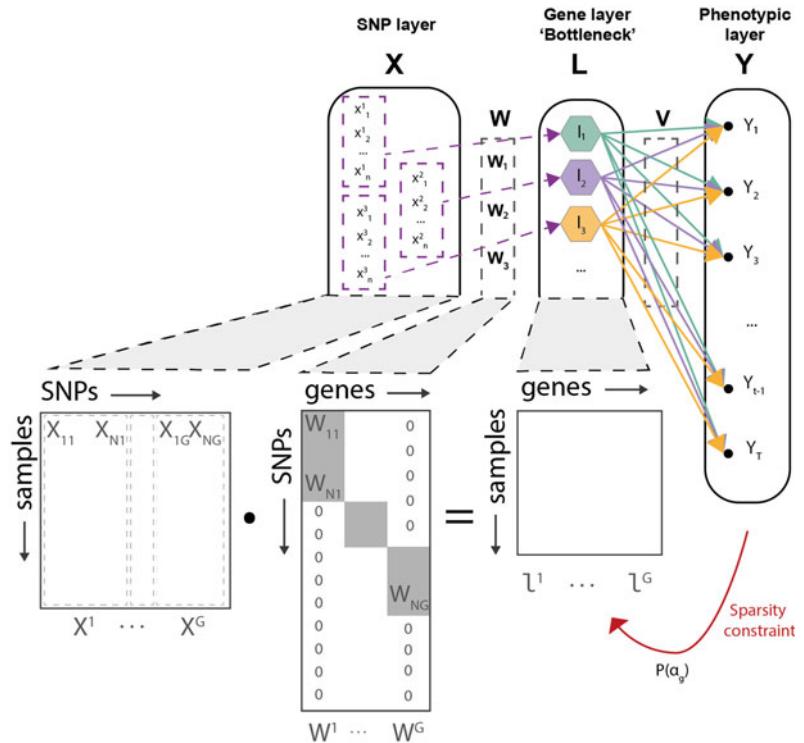


Fig. 6 Illustration of G2PSR SNP-gene grouping constraint and overall neural network architecture

parametrized as $q(\mathbf{W}^T)$, such that each element of the input layer is defined as $W_i^T \sim \mathcal{N}(\mu_i^T; \alpha_g \cdot \mu_i^{T2})$ [57], where the parameter α_g is optimized to quantify the common uncertainty associated with the ensemble of SNPs contributing to the gene g .

5 Conclusions

This chapter presented an overview of basic notions and tools for multimodal analysis. The set of frameworks introduced here represents an ideal starting ground for more complex analysis, either based on linear multivariate methods [58, 59] or on neural network architectures, extending the modeling capabilities to account for highly heterogeneous information, such multi-organ data [46], text information, and data from electronic health records [60, 61].

Acknowledgements

This work was supported by the French government, through the 3IA Côte d’Azur Investments in the Future project managed by the National Research Agency (ANR) (ANR-19-P3IA-0002).

References

1. Civelek M, Lusis AJ (2014) Systems genetics approaches to understand complex traits. *Nat Rev Gen* 15(1):34–48. <https://doi.org/10.1038/nrg3575>
2. Liu S, Cai W, Liu S, Zhang F, Fulham M, Feng D, Pujol S, Kikinis R (2015) Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders. *Brain Inform* 2(3):167–180. <https://doi.org/10.1007/s40708-015-0019-x>
3. Shen L, Thompson PM (2020) Brain imaging genomics: Integrated analysis and machine learning. *Proc IEEE Inst Electr Electron Eng* 108(1):125–162. <https://doi.org/10.1109/JPROC.2019.2947272>
4. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet* 101(1):5–22. <https://doi.org/10.1016/j.ajhg.2017.06.005>
5. Lahat D, Adali T, Jutten C (2014) Challenges in multimodal data fusion. In: EUSIPCO 2014—22th European signal processing conference, Lisbonne, Portugal, pp 101–105. <https://hal.archives-ouvertes.fr/hal-01062366>
6. Menon BK, Campbell BC, Levi C, Goyal M (2015) Role of imaging in current acute ischemic stroke workflow for endovascular therapy. *Stroke* 46(6):1453–1461. <https://doi.org/10.1161/STROKEAHA.115.009160>
7. Zameer S, Siddiqui AS, Riaz R (2021) Multimodality imaging in acute ischemic stroke. *Curr Med Imaging* 17(5):567–577
8. Liu X, Lai Y, Wang X, Hao C, Chen L, Zhou Z, Yu X, Hong N (2013) A combined DTI and structural MRI study in medicated-naïve chronic schizophrenia. *Magn Reson Imaging* 32(1):1–8
9. Rastogi S, Lee C, Salamon N (2008) Neuroimaging in pediatric epilepsy: a multimodality approach. *Radiographics* 28(4):1079–1095
10. Abela E, Rummel C, Hauf M, Weisstanner C, Schindler K, Wiest R (2014) Neuroimaging of epilepsy: lesions, networks, oscillations. *Clin Neuroradiol* 24(1):5–15
11. Fernández S, Donaire A, Serès E, Setoain X, Bargalló N, Falcén C, Samartí F, Maestro I, Rumià J, Pintor L, Boget T, Aparicio J, Carreño M (2015) PET/MRI and PET/MRI/SISCOM coregistration in the presurgical evaluation of refractory focal epilepsy. *Epilepsy Research* 111:1–9. <https://doi.org/10.1016/j.eplepsyres.2014.12.011>
12. Hong SB, Zalesky A, Fornito A, Park S, Yang YH, Park MH, Song IC, Sohn CH, Shin MS, Kim BN, Cho SC, Han DH, Cheong JH, Kim JW (2014) Connectomic disturbances in attention-deficit/hyperactivity disorder: a whole-brain tractography analysis. *Biol Psychiatry* 76(8):656–663
13. Mueller S, Keeser D, Samson AC, Kirsch V, Blautzik J, Grothe M, Erat O, Hegenloh M, Coates U, Reiser MF, Hennig-Fast K, Meindl T (2013) Convergent findings of altered functional and structural brain connectivity in individuals with high functioning autism: a multimodal mri study. *PLOS ONE* 8(6):1–11. <https://doi.org/10.1371/journal.pone.0067329>
14. Lorenzi M, Altmann A, Gutman B, Wray S, Arber C, Hibar DP, Jahanshad N, Schott JM, Alexander DC, Thompson PM, Ourselin S, null null (2018) Susceptibility of brain atrophy to *TRIB3* in Alzheimer's disease, evidence from functional prioritization in imaging genetics. *Proc Natl Acad Sci* 115(12):3162–3167. <https://doi.org/10.1073/pnas.1706100115>
15. Kim M, Kim J, Lee SH, Park H (2017) Imaging genetics approach to Parkinson's disease and its correlation with clinical score. *Sci Rep* 7(1):46700. <https://doi.org/10.1038/srep46700>
16. Martins D, Giacomet A, Williams SC, Turkheimer F, Dipasquale O, Veronese M, Group PTW, et al. (2021) Imaging transcriptomics: convergent cellular, transcriptomic, and molecular neuroimaging signatures in the healthy adult human brain. *Cell Rep* 37(13):110173
17. Schrouff J, Rosa MJ, Rondina JM, Marquand AF, Chu C, Ashburner J, Phillips C, Richiardi J, Mourão-Miranda J (2013) PRONTO: pattern recognition for neuroimaging toolbox. *Neuroinformatics* 11(3):319–337
18. Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, Zhang H, Sun YV (2009) Machine learning in genome-wide association studies. *Genetic Epidemiol* 33(S1):S51–S57
19. Liu J, Calhoun VD (2014) A review of multivariate analyses in imaging genetics. *Front Neuroinform* 8:29
20. Lorenzi M, Altmann A, Gutman B, Wray S, Arber C, Hibar DP, Jahanshad N, Schott JM, Alexander DC, Thompson PM, Ourselin S (2018) Susceptibility of brain atrophy to *trib3* in Alzheimer's disease, evidence from

- functional prioritization in imaging genetics. *Proc Natl Acad Sci* 115(12):3162–3167. <https://doi.org/10.1073/pnas.1706100115>
21. Shashanka M, Raj B, Smaragdis P (2007) Sparse overcomplete latent variable decomposition of counts data. In: Advances in neural information processing systems, vol 20
 22. Anandkumar A, Ge R, Janzamin M (2015) Learning overcomplete latent variable models through tensor methods. In: Conference on learning theory, PMLR, pp 36–112
 23. Antelman L, Ayache N, Robert P, Lorenzi M (2019) Sparse multi-channel variational auto-encoder for the joint analysis of heterogeneous data. In: International conference on machine learning, PMLR, pp 302–311
 24. Hotelling H (1936) Relations between two sets of variates. *Biometrika* 28(3/4):321
 25. Liu J, Calhoun V (2014) A review of multivariate analyses in imaging genetics. *Front Neuroinform* 8:29. <https://doi.org/10.3389/fninf.2014.00029>
 26. Kettenring JR (1971) Canonical analysis of several sets of variables. *Biometrika* 58(3): 433–451. <https://doi.org/10.1093/biomet/58.3.433>
 27. Luo Y, Tao D, Ramamohanarao K, Xu C, Wen Y (2015) Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE Trans Knowl Data Eng* 27(11):3111–3124. <https://doi.org/10.1109/TKDE.2015.2445757>
 28. Huang SY, Lee MH, Hsiao CK (2009) Non-linear measures of association with kernel canonical correlation analysis and applications. *J Stat Plan Inference* 139(7):2162–2174. <https://doi.org/10.1016/j.jspi.2008.10.011>
 29. Andrew G, Arora R, Bilmes J, Livescu K (2013) Deep canonical correlation analysis. In: Dasgupta S, McAllester D (eds) Proceedings of the 30th international conference on machine learning, PMLR, Atlanta, Georgia, USA, Proceedings of Machine Learning Research, vol 28, pp 1247–1255. <https://proceedings.mlr.press/v28/andrew13.html>
 30. McIntosh A, Bookstein F, Haxby JV, Grady C (1996) Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage* 3(3):143–157
 31. Worsley KJ (1997) An overview and some new developments in the statistical analysis of pet and fmri data. *Hum Brain Mapp* 5(4):254–258
 32. De Bie T, Cristianini N, Rosipal R (2005) Eigenproblems in pattern recognition. In: Handbook of geometric computing, pp 129–167
 33. Bach F, Jordan M (2003) Kernel independent component analysis. *J Mach Learn Res* 3:1–48. <https://doi.org/10.1162/153244303768966085>
 34. Theodoridis S, Koutroumbas K (2008) Pattern recognition, 4th edn. Academic Press, New York
 35. Wold H (1975) Path models with latent variables: the nipa approach. In: Quantitative sociology. Elsevier, Amsterdam, pp 307–357
 36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
 37. Tenenhaus M (1999) L'approche pls. *Revue de statistique appliquée* 47(2):5–40
 38. Vidaurre D, van Gerven MA, Bielza C, Larrañaga P, Heskes T (2013) Bayesian sparse partial least squares. *Neural Comput* 25(12): 3318–3339
 39. Klami A, Virtanen S, Kaski S (2013) Bayesian canonical correlation analysis. *J Mach Learn Res* 14(4):965–1003
 40. Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. *J R Stat Soc Series B (Statistical Methodology)* 61(3): 611–622
 41. Balelli I, Silva S, Lorenzi M (2021) A probabilistic framework for modeling the variability across federated datasets of heterogeneous multi-view observations. In: Information processing in medical imaging: proceedings of the...conference.
 42. Kingma DP, Welling M (2014) Auto-Encoding Variational Bayes. In: Proc. 2nd Int. Conf. Learn. Represent. (ICLR2014) 1312.6114
 43. Rezende DJ, Mohamed S, Wierstra D (2014) Stochastic backpropagation and approximate inference in deep generative models. In: International conference on machine learning, PMLR, pp 1278–1286
 44. Kim Y, Wiseman S, Miller A, Sontag D, Rush A (2018) Semi-amortized variational autoencoders. In: International conference on machine learning, PMLR, pp 2678–2687
 45. Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: a review for statisticians. *J Am Stat Assoc* 112(518):859–877
 46. Diaz-Pinto A, Ravikumar N, Attar R, Suinesiaputra A, Zhao Y, Levelt E, Dall'Armellina E, Lorenzi M, Chen Q, Keenan TD et al (2022) Predicting myocardial infarction through retinal scans and minimal personal information. *Nat Mach Intell* 4:55–61

47. Nocedal J, Wright S (2006) Numerical optimization. Springer nature, pp 1–664. Springer series in operations research and financial engineering
48. Wang W, Lee H, Livescu K (2016) Deep variational canonical correlation analysis. <http://arxiv.org/abs/1610.03454>
49. Hafkemeijer A, Altmann-Schneider I, Oleksik AM, van de Wiel L, Middelkoop HA, van Buchem MA, van der Grond J, Rombouts SA (2013) Increased functional connectivity and brain atrophy in elderly with subjective memory complaints. *Brain Connectivity* 3(4): 353–362
50. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B (Statistical Methodology)* 68(1):49–67
51. Zhang Y, Xu Z, Shen X, Pan W, Initiative ADN (2014) Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage* 96:309–325. <https://doi.org/10.1016/j.neuroimage.2014.03.061>
52. Hibar DP, Stein JL, Kohannim O, Jahanshad N, Saykin AJ, Shen L, Kim S, Pankratz N, Foroud T, Huentelman MJ, Potkin SG, Jack Jr CR, Weiner MW, Toga AW, Thompson PM, Initiative ADN (2011) Voxel-wise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. *NeuroImage* 56(4): 1875–1891. <https://doi.org/10.1016/j.neuroimage.2011.03.077>
53. Ge T, Feng J, Hibar DP, Thompson PM, Nichols TE (2012) Increasing power for voxel-wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures. *NeuroImage* 63:858–873
54. Schmidt W, Kraaijveld M, Duin R (1992) Feedforward neural networks with random weights. In: Proceedings of the 11th IAPR international conference on pattern recognition. Vol. II. Conference B: pattern recognition methodology and systems, pp 1–4. <https://doi.org/10.1109/ICPR.1992.201708>
55. Deprez M, Moreira J, Sermesant M, Lorenzi M (2022) Decoding genetic markers of multiple phenotypic layers through biologically constrained genome-to-phenome Bayesian sparse regression. *Front Mol Med*. <https://doi.org/10.3389/fmmed.2022.830956>
56. Molchanov D, Ashukha A, Vetrov D (2017) Variational dropout sparsifies deep neural networks. arXiv 1701.05369
57. Kingma DP, Welling M (2014) Auto-encoding variational bayes. CoRR abs/1312.6114
58. Pearson GD, Liu J, Calhoun VD (2015) An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders. *Front Genetics* 6:276
59. Le Floch É, Guillemot V, Frouin V, Pinel P, Lalanne C, Trinchera L, Tenenhaus A, Moreno A, Zilbovicius M, Bourgeron T et al (2012) Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. *NeuroImage* 63(1):11–24
60. Rodin I, Fedulova I, Shelmanov A, Dylov DV (2019) Multitask and multimodal neural network model for interpretable analysis of x-ray images. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, pp 1601–1604
61. Huang SC, Pareek A, Zamani R, Banerjee I, Lungren MP (2020) Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. *Sci Rep* 10(1):1–9

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made. The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part IV

Validation and Datasets



Chapter 20

Evaluating Machine Learning Models and Their Diagnostic Value

Gael Varoquaux and Olivier Colliot

Abstract

This chapter describes model validation, a crucial part of machine learning whether it is to select the best model or to assess performance of a given model. We start by detailing the main performance metrics for different tasks (classification, regression), and how they may be interpreted, including in the face of class imbalance, varying prevalence, or asymmetric cost–benefit trade-offs. We then explain how to estimate these metrics in an unbiased manner using training, validation, and test sets. We describe cross-validation procedures—to use a larger part of the data for both training and testing—and the dangers of data leakage—optimism bias due to training data contaminating the test set. Finally, we discuss how to obtain confidence intervals of performance metrics, distinguishing two situations: internal validation or evaluation of learning algorithms and external validation or evaluation of resulting prediction models.

Key words Validation, Performance metrics, Cross-validation, Data leakage, External validation

1 Introduction

A machine learning (ML) model is validated by evaluating its prediction performance. Ideally, this evaluation should be representative of how the model would perform when deployed in a real-life setting. This is an ambitious goal that goes beyond the settings of academic research. Indeed, a perfect validation would probe robustness to any possible variation of the input data that may include different acquisition devices and protocols, different practices that vary from one country to another, from one hospital to another, and even from one physician to another. A less ambitious goal for validation is to provide an unbiased estimate of the model performance on new—never before seen—data similar to that used for training (but not the same data!). By similar, we mean data that have similar clinical or sociodemographic characteristics and that have been acquired using similar devices and protocols. To go beyond such *internal validity*, external validation would evaluate

generalization to data from different sources (for example, another dataset, data from another hospital).

This chapter addresses the following questions. How to quantify the performance of the model? This will lead us to present, in Subheading 2, different performance metrics that are adequate for different ML tasks (classification, regression, ...). How to estimate these performance metrics? This will lead to the presentation of different validation strategies (Subheading 3). We will also explain how to derive confidence intervals for the estimated performance metrics, drawing the distinction between evaluating a learning algorithm or a resulting prediction model. We will present various caveats that pertain to the use of performance metrics on medical data as well as to data leakage, which can be particularly insidious.

2 Performance Metrics

Metrics allow to quantify the performance of an ML model. In this section, we describe metrics for classification and regression tasks. Other tasks (segmentation, generation, detection,...) can use some of these but will often require other metrics that are specific to these tasks. The reader may refer to Chap. 13 for metrics dedicated to segmentation and to Subheading 6 of Chap. 23 for metrics dedicated to segmentation, classification, and detection.

2.1 Metrics for Classification

2.1.1 Binary Classification

For classification tasks, the results can be summarized in a matrix called the confusion matrix (Fig. 1). For binary classification, the confusion matrix divides the test samples into four categories, depending on their true and predicted labels:

		True label	
		Positive	Negative
Predicted label	Positive	$D+$	$D-$
	Negative	$T+$	$T-$
	$T+$	TP	FP
	$T-$	FN	TN

Fig. 1 Confusion matrix. The confusion matrix represents the results of a classification task. In the case of binary classification (two classes), it divides the test samples into four categories, depending on their true (e.g., disease status, D) and predicted (test output, T) labels: true positives (TP), true negatives (TN), false positives (FP), false negatives (FN)

- **True Positives (TP):** Samples for which the true and predicted labels are both 1. Example: The patient has cancer (1), and the model classifies this sample as cancer (1).
- **True Negatives (TN):** Samples for which the true and predicted labels are both 0. Example: The patient does not have cancer (0), and the model classifies this sample as non-cancer (0).
- **False Positives (FP):** Samples for which the true label is 0 and the predicted label is 1. Example: The patient does not have cancer (0), and the model classifies this sample as cancer (1).
- **False Negatives (FN):** Samples for which the true label is 1 and the predicted label is 0. Example: The patient has cancer (1), and the model classifies this sample as non-cancer (0).

Are false positives and false negatives equally problematic? This depends on the application. For instance, consider the case of detecting brain tumors. For a screening application, detected positive cases would then be subsequently reviewed by a human expert, and one can thus consider that false negatives (missed brain tumor) lead to more dramatic consequences than false positives. On the opposite, if a detected tumor leads the patient to be sent to brain surgery without complementary exam, false positives are problematic and brain surgery is not a benign operation. For automatic volumetry from magnetic resonance images (MRI), one could argue that false positives and false negatives are equally problematic.

Box 1: Performance Metrics for Binary Classification

Basic metrics

T denotes *test*: classifier output; D denotes *diseased* status.

- **Sensitivity (also called recall):** A fraction of positive samples actually retrieved.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{Estimates } P(T+ | D+).$$

- **Specificity:** A fraction of negative samples actually classified as negative.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad \text{Estimates } P(T- | D-).$$

- **Positive predictive value (PPV, also called precision):** A fraction of the positively classified samples that are indeed positive.

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{Estimates } P(D+ | T+).$$

- **Negative predictive value (NPV):** A fraction of the negatively classified samples that are indeed negative.

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad \text{Estimates } P(D- | T-).$$

(continued)

Box 1 (continued)
Summary metrics

- **Accuracy:** A fraction of the samples correctly classified.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}.$$
- **Balanced accuracy (BA):** Accuracy metric that accounts for unbalanced samples.

$$\text{BA} = \frac{\text{Sensitivity} + \text{Specificity}}{2}.$$
- **F_1 score:** Harmonic mean of PPV (precision) and sensitivity (recall).

$$F_1 = \frac{2}{\frac{1}{\text{PPV}} + \frac{1}{\text{Sensitivity}}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}.$$
- **Matthews correlation coefficient (MCC):** MCC=1 for perfect classification, MCC=0 for random classification, MCC=-1 for perfectly wrong classification.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}}.$$
- **Markedness** = $\frac{\text{TP}}{\text{TP} + \text{FP}} - \frac{\text{FP}}{\text{FP} + \text{TN}} = \text{PPV} + \text{NPV} - 1$.
- **Area under the receiver operating characteristic curve (ROC AUC).**
- **Area under the precision-recall curve (PR AUC, also called average precision).**

Multiple performance metrics can be derived from the confusion matrix, all easily computed using `sklearn.metrics` from scikit-learn [1]. They are summarized in Box 1. One can distinguish between basic metrics that only focus on false positives or false negatives and summary metrics that aim at providing an overview of the performance with a single metric.

The performance of a classifier is characterized by pairs of basic metrics: either sensitivity and specificity, or PPV and NPV, which characterize respectively the probability of the test given the diseased status or vice versa (see Box 1). Note that each basic metric characterizes only the behavior of the classifier on the positive class ($D+$) or the negative class ($D-$); thus measuring both sensitivity *and* specificity and PPV *and* NPV is important. Indeed, a classifier always reporting a positive prediction would have a perfect sensitivity, but a disastrous specificity.

It is convenient to use summary metrics that provide a more global assessment of the performance, for instance, to select a “best” model. However, as we will see, summary metrics, when used in isolation, can lead to erroneous conclusions. The most widely used summary metric is arguably accuracy. Its main advantage is a natural interpretation: the proportion of correctly classified samples. However, it is misleading when the data are imbalanced. Let us for

instance consider a dataset with 10 cancer samples and 990 non-cancer samples. A trivial majority classifier that decides that cancer does not exist achieves 99% accuracy. Balanced accuracy helps for imbalanced samples. However, balanced accuracy also comes with its loopholes. Indeed, a high balanced accuracy does not always mean that individuals classified as diseased are likely to be so. Let us consider a diagnostic test for a disease that has a sensitivity of 99% and a specificity of 90% (and thus a balanced accuracy of 94.5%). Suppose that a given person takes the test and that the test is positive. At this point, we do not have enough information to compute the probability that the person actually has the disease.

The probability that the person has the disease is given by the PPV, related to the sensitivity and the specificity by Bayes' rule:

$$P(D+ | T+) = \frac{\text{sensitivity} \times \text{prevalence}}{(1 - \text{specificity}) \times (1 - \text{prevalence}) + \text{sensitivity} \times \text{prevalence}}.$$

Diseased
↑
Test positive

Bayes' rule thus shows that we must account for the prevalence: the proportion of the people with the disease in the target population, the population in which the test is intended to be applied. The target population can be the general population for a screening test. It could be the population of people with memory complaints for a test aiming to diagnose Alzheimer's disease. Now, suppose that the prevalence is low, which will often be the case for a screening test in the general population. For instance, prevalence = 0.001. This leads to $P(D+ | T+) = 0.0098 \approx 1\%$. So, if the test is positive, there is only 1% chance that the patient has the disease. Even though our classifier has seemingly good sensitivity, specificity, and balanced accuracy, it is not very informative on the general population. The PPV and NPV readily give the information of interest: $P(D+ | T+)$ and $P(D- | T-)$. However, they are not natural metrics to report a classifier's performance because, unlike sensitivity and specificity, they are not intrinsic to the test (in other words the trained ML model) but also depend on the prevalence and thus on the target population (Fig. 2).

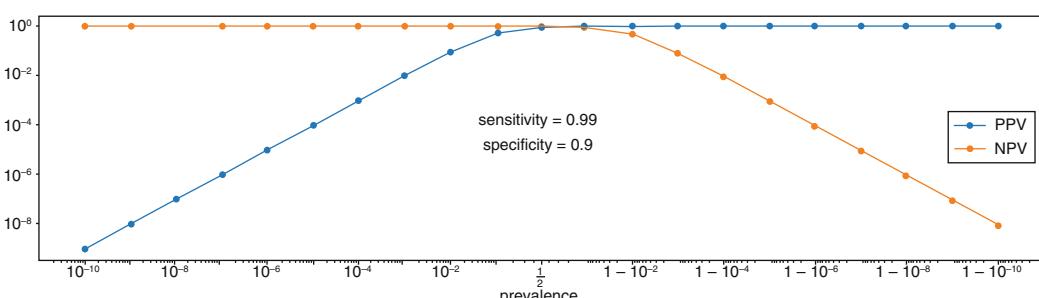


Fig. 2 NPV and PPV as functions of prevalence when the sensitivity and the specificity are fixed (image courtesy of Johann Faouzi)

Summary Metrics for Low Prevalence

The F_1 score is another summary metric, built as the harmonic mean of the sensitivity (recall) and PPV (precision). It is popular in machine learning but, as we will see, it also has substantial drawbacks. Note that it is equal to the Dice coefficient used for segmentation. Given that it builds on the PPV rather than the specificity to characterize retrieval, it accounts slightly better for prevalence. In our example, the F_1 score would have been low. The F_1 score can nevertheless be misleading if the prevalence is high. In such a case, one can have high values for sensitivity, specificity, PPV, F_1 score but a low NPV. A solution can be to exchange the two classes. The F_1 score becomes informative again. Those shortcomings are fundamental, as the F_1 score is completely blind to the number of true negatives, TNs. This is probably one of the reasons why it is a popular metric for segmentation (usually called Dice rather than F_1) as in this task TN is almost meaningless (TN can be made arbitrarily large by just changing the field of view of the image). In addition, this metric has no simple link to the probabilities of interest, even more so after switching classes.

Another option is to use Matthews Correlation Coefficient (MCC). The MCC makes full use of the confusion matrix and can remain informative even when prevalence is very low or very high. However, its interpretation may be less intuitive than that of the other metrics. Finally, *markedness* [2] is a seldom known summary metric that deals well with low-prevalence situations as it is built from the PPV and NPV (Box 1). Its drawback is that it is as much related to the population under study as to the classifier.

As we have seen, it is important to distinguish metrics that are intrinsic characteristics of the classifier (sensitivity, specificity, balanced accuracy) from those that are dependent on the target population and in particular of its prevalence (PPV, NPV, MCC, markedness). The former are independent of the situation in which the model is going to be used. The latter inform on the probability of the condition (the output label) given the output of the classifier, but they depend on the operational situation and, in particular, on the prevalence. The prevalence can be variable (for instance, the prevalence of an infectious disease will be variable across time, and the prevalence of a neurodegenerative disease will depend on the age of the target population), and a given classifier may be intended to be applied in various situations. This is why the intrinsic characteristics (sensitivity and specificity) need to be judged according to the different intended uses of the classifier (e.g., a specificity of 90% may be considered excellent for some applications, while it would be considered unacceptable if the intended use is in a low-prevalence situation).

Metrics for Shifts in Prevalence

Odds enable designing metrics that characterize the classifier but are adapted to target populations with a low prevalence. Odds are defined as the ratio between the probability that an event occurs and the probability this event does not occur: $\mathcal{O}(\alpha) = \frac{P(\alpha)}{1-P(\alpha)}$. Ratios between odds can be invariant to the sampling frequency (or prevalence) of α —see Appendix “Odds Ratio and Diagnostic Tests Evaluation” for an introduction to odds and their important properties. For this reason, they are often used in epidemiology. A classifier can be characterized by the ratio between the pre-test and post-test odds, often called the *positive likelihood ratio*: $LR+ = \frac{\mathcal{O}(D+|T+)}{\mathcal{O}(D+)} = \frac{\text{sensitivity}}{1-\text{specificity}}$. This quantity depends only on sensitivity and specificity, properties of the classifier only, and not of the prevalence on the study population. Yet, given a target population, post-test odds can easily be obtained by multiplying $LR+$ by pre-test odds, itself given by prevalence: $\mathcal{O}(D+) = \frac{\text{prevalence}}{1-\text{prevalence}}$. The larger the $LR+$, the more useful the classifier and a classifier with $LR+ = 1$ or less brings no additional information on the likelihood of the disease. An equivalent to $LR+$ characterizes the negative class: controlling on “ T^- ” instead of “ T^+ ” gives the *negative likelihood ratio*: $LR- = \frac{1-\text{sensitivity}}{\text{specificity}}$; and low values of $LR-$ (below 1) denote more useful predictions. These metrics, $LR+$ and $LR-$, are very useful in a situation common in biomedical settings where the only data available to learn and evaluate a classifier are *study* population with nearly balanced classes, such as a case-control study, while the target application—the general population—is one with a different prevalence (e.g., a very low prevalence) or when the intended use considers variable prevalences.

Multi-threshold Metrics

Many classification algorithms output a continuous value that is then thresholded to get a binary label. When the output is a probability, one often simply uses a threshold of 0.5. However, there are cases where one is interested to study the performance for varying thresholds on the output. The two main tools for that purpose are the receiver operating characteristic (ROC) curve and the precision-recall (PR) curve. The ROC curve plots the Sensitivity as a function of 1 – Specificity (Fig. 3). It can be again summarized with a single value: the area under the ROC curve (ROC AUC). The ROC AUC has a probabilistic interpretation: it is the probability that a positive sample has a higher classification score (as positive) than a negative sample. A perfect classification corresponds to an ROC AUC of 1 and a random classification to an ROC AUC of 0.5. While chance remains 0.5 whatever the class imbalance, the ROC curve becomes less interesting for highly imbalanced classes, because a seemingly small difference on specificity or sensitivity may make a large difference to the application, but not change much the ROC curve. For this reason, it is often

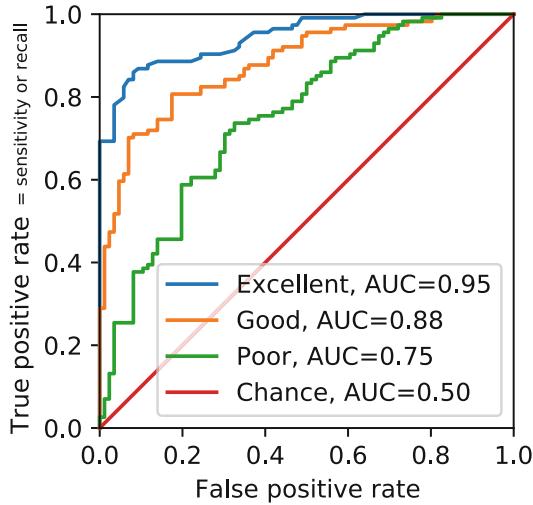


Fig. 3 ROC curve for different classifiers. AUC denotes the area under the curve, typically used to extract a number summarizing the ROC curve

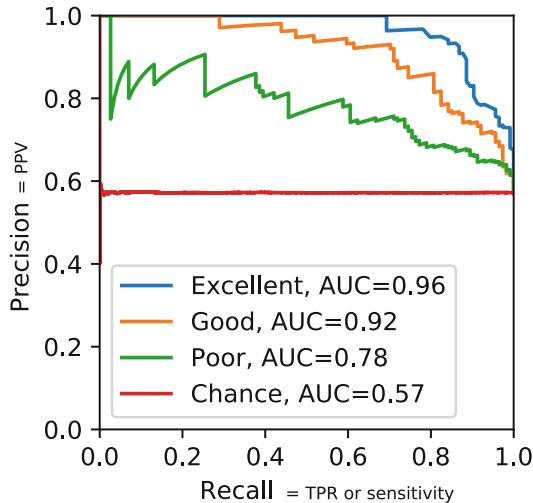


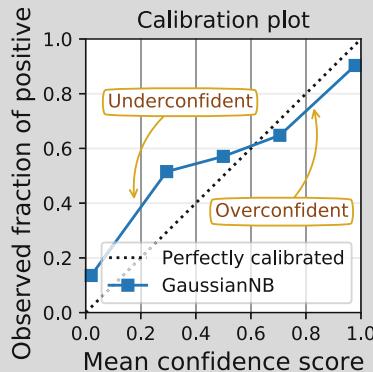
Fig. 4 Precision–recall curve for different classifiers. AUC denotes the area under the curve, often called *average precision* here. Note that the chance level depends on the class imbalance (or prevalence), here 0.57

complemented with the precision–recall (PR) curve that focuses on the minority class. The PR curve plots the Precision (also called PPV) as a function of Recall (also called sensitivity) (Fig. 4). It can also be summarized using a single measure: the PR AUC, also called average precision. As for the ROC AUC, a perfect classification corresponds to a value of 1. However, unlike for ROC AUC, a dummy classification does not necessarily lead to a value of 0.5. It depends on the prevalence.

Box 2: Assessing Confidence Scores and Calibration

Expected calibration error (ECE): *average classifier error*

It is computed by considering K bins of confidence scores and comparing the observed fraction of positives to the mean confidence score. The ECE itself is then the average over the bins: $ECE = \sum_{i=1}^K P(i) \cdot |f_i - s_i|$, where f_i is the observed fraction of positive instances in bin i , s_i is the mean of classifier scores for the instances in bin i , and $P(i)$ is the fraction of all instances that fall into bin i [3].



Example for a Gaussian Naive Bayes classifier (GaussianNB).

Metrics on individual probabilities: *error on $P(y|X)$*

$$\text{Brier score} = \sum_i (\hat{s}_i - y_i)^2$$

Observed (binary) label
Confidence score

Minimal for $\hat{S} = P(y|X)$

$$\text{Brier skill score} = 1 - \frac{\text{Brier}(\hat{s}, y)}{\text{Brier}(\bar{y}, y)}$$

Class prevalence

A value of 1 means a perfect prediction, while a value of 0 means that the confidence scores are not more informative than the class prevalence.

Confidence Scores and Calibration

It can be useful to interpret a non-thresholded classifier score as a confidence score or a probability, for instance, to balance cost and benefits when the prediction is used to decide on an intervention [4]. But a continuous score by itself does not warrant such interpretation: a classifier may be over-confident, under-confident, or have uneven scores over the population, even for good binary decisions. Two types of metrics, detailed in Box 2, are useful to

evaluate continuous outputs as probabilities: the *expected calibration error* (ECE) and the *Brier score*. The ECE measures whether, on samples predicted with a score s , the error rate is indeed s , in which case the classifier is said to be calibrated. The *Brier score* is minimal when the classifier score is the true probability of the class given the data for an individual, for instance, the probability of the presence of a tumor given the image. These two notions are similar, but it is important to understand that ECE controls average error rates, while Brier score controls individual probabilities, which is much more stringent and more useful to the practitioner [5]. Accurate probabilities of individual predictions can be used for optimal decision-making, *e.g.*, opting for brain surgery only for individuals for which a diagnostic model predicts cancer with high confidence.

A given value of ECE is easy to interpret, as it qualifies probabilities mostly independently of prediction performance. On the other hand, the Brier score accounts for both the quality of probabilities and corresponding binary decisions as a low Brier score captures the ability to give good probabilistic prediction of the output. For any classification problem, there exist many classifiers with 0 expected calibration errors, including some with very poor predictions. On the other hand, even the best possible prediction has a non-zero Brier score, unless the output is a deterministic function of the data. The Brier skill score, a variant of the Brier score, is often used to assess how far a predictor is from the best possible prediction, more independent of the intrinsic uncertainty in the data. The Brier skill score is a rescaled version of the Brier score taking as a reference a reasonable baseline: 1 is a perfect prediction, while negative values mean predictions worse than guessing from class prevalence.

To Conclude

When assessing a classifier:

- Always look at all the individual metrics: false positives and false negatives are seldom equivalent. Understand the medical problem to know the right trade-off [4].
- Never trust a single summary metric (accuracy, balanced accuracy, ROC AUC, ...).
- Consider the prevalence in your target population. It may be that the prevalence in your testing sample is not representative of that of the target population. In that case, aside from LR+ and LR-, performance metrics computed from the testing sample will not be representative of those in the target population.

2.1.2 Multi-class Classification

When there are multiple classes to distinguish, the main difference with two-class classification is that the problem can no longer be separated into a positive class (typically individuals with the medical condition of interest) and a negative class (individuals without). As a consequence, sensitivity and specificity no longer have a meaning

for the whole data, nor do F_1 score, or the ROC or precision–recall curves. Accuracy is still defined and easy to compute, but still suffers from its common drawbacks, in particular that it may not be straightforward to interpret in the face of class imbalance.

A classic approach is to aggregate metrics for binary settings considering successively each class as the positive instances and all the others as the negatives, in a form of “one versus all.” There are different approaches to averaging the results for each class. *Macro*-averaging computes the metric, for instance, the ROC AUC, for each class, and then averages the results. One drawback is that it may put too much emphasis on classes that are more infrequent. *Weighted* or *micro*-averaging combines the results of the different classes weighed by the number of instances of each class. The difference between the two is that *weighted* averaging computes the average of the metric weighted by the number of true instances for each class, while *micro*-averaging computes the metric by adding the number of TPs (resp., TNs, FPs, FNs) across all classes.

Inspecting the confusion matrix extended to multi-class settings gives an interesting tool to understand errors: it displays how many times a given true class is predicted as another (Fig. 5). A perfect prediction has non-zero entries only on the diagonal. The confusion matrix may be interesting to reveal which classes are commonly confused, as its name suggests. In our example, instances that are actually of class C2 are often predicted as of class C3.

Multilabel Classification

Multilabel settings are when the multiple classes are not mutually exclusive: for instance, if an individual can have multiple pathologies. The problem is then to detect the presence or absence of each label for an individual. In terms of evaluation, multilabel settings can be understood as several binary classification problems, and thus the corresponding metrics can be used on each label. As in the multi-class settings, there are different ways to average the results for each label—macro, micro—that put more or less emphasis on the rare labels.

		Predicted		
		C1	C2	C3
True	C1	133	0	0
	C2	0	107	36
	C3	0	0	92

Fig. 5 Multi-class confusion matrix, for a 3-class problem, C1, C2, C3. Each entry gives the number of instances predicted of a given class, knowing the actual class. A perfect prediction would give non-zero entries only on the diagonal

2.2 Metrics for Regression

In regression settings, the outcome to predict y is continuous, for instance, an individual’s age, cognitive scores, or glucose level. Corresponding error metrics gauge how far the prediction \hat{y} is from the observed y .

R2 Score. The go-to metric here is typically the R^2 score, sometimes called explained variance—however, the term R^2 score should be preferred, as some authors define explained variance as ignoring bias. Mathematically, the R^2 score is the fraction of variance of the outcome y explained by the prediction \hat{y} , relative to the variance explained by the mean \bar{y} on the test set:

$$R^2 = 1 - \frac{\text{SS}(y - \hat{y})}{\text{SS}(y - \bar{y})},$$

where SS is the sum of squares on the test data. A strong benefit of this metric is that it comes with a natural scale: an R^2 of 1 implies perfect prediction, while an R^2 of zero implies a trivial and not very useful prediction. Note that chance-level predictions (as obtained for instance by learning on permuted y) yield slightly negative predictions: indeed, even when the data do not support a prediction of y —as in chance settings—it is impossible to estimate the mean y perfectly and predictions will be worse than the actual mean. In this respect, the R^2 score has a different behavior in machine learning settings compared to inferential statistics settings not focused on prediction: *in-sample* (for inferential statistics) versus *out-of-sample* settings (for machine learning). Indeed, when the mean of y is computed on the same data as the model, the R^2 score is positive and is the square of the correlation between y and \hat{y} . This is not the case in predictive settings, and the correlation between y and \hat{y} should not be used to judge the quality of a prediction [6], because it discards errors on the mean and the scale of the prediction, which are important in practice.

Absolute Error Measures. Reporting only the R^2 score is not sufficient to characterize well a predictive model. Indeed, the R^2 score depends on the variance of the outcome y in the study population and thus does not enable comparing predictive models on different samples. For this purpose, it is important to report also an absolute error measure. The root mean square error (RMSE) and the mean absolute error (MAE) are two of such measures that give an error in the scale of the outcome: if the outcome y is an age in years, the error is also in years. The mean absolute error is easier to interpret. Compared to the root mean square error, the mean absolute error will put much less weight on some rare large deviations. For instance, consider the following prediction error (on 11 observations):

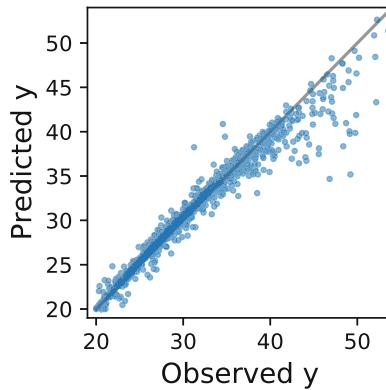


Fig. 6 Visualizing prediction errors—plotting the predicted outcome as a function of the observed one enables to detect structure in the error beyond summary metric. Here the error increases for large values of y , for which there is also a systematic undershoot

$$\begin{aligned} \text{error} &= [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 100] \\ \text{MAE} &= 10 \qquad \qquad \qquad \text{RMSE} \approx 30.17. \end{aligned}$$

Note that if the error was uniformly equal to the same value (10, for instance), both measures would give the same result.

Assessing the Distribution of Errors. The difference between the mean absolute error and the root mean square error arises from the fact that both measures account differently for the tails of the distribution of errors. It is often useful to visualize these errors, to understand how they are structured. Figure 6 shows such visualization: predicted y as a function of observed y . It reveals that for large values of y , the predictive model has a larger prediction error, but also that it tends to undershoot: predict a value that underestimates the observed value. This aspect of the prediction error is not well captured by the summary metrics because there are comparatively much less observations with large y .

Concluding Remarks on Performance Metrics. Whether it is in regression or in classification, a single metric is not enough to capture all aspects of prediction performance that are important for applications. Heterogeneity of the error, as we have just seen in our last example, can be present not only as a function of prediction target, but of any aspect of the problem, for instance, the sex of the individuals. Problems related to *fairness*, where some groups (e.g., demographic, geographic, socioeconomic groups) suffer more errors than others, can lead to loss of trust or amplification of inequalities [7]. For these reasons, it may be important to also report error metrics on relevant subgroups, following common medical research practice of stratification.

3 Evaluation Strategies

The previous section detailed metrics for assessing the performance of a ML model. We now focus on how to estimate the expected prediction performance of the model with these metrics. Importantly, we draw the difference between evaluating a learning procedure, or learner, and a learned model. While these two questions are often conflated in the literature, the first one must account for uncontrolled fluctuations in the learning procedure, while the second one controls a given model on a target external population. The first question is typically of interest to the methods researcher, to conclude on learning procedures, while the second is central to the medical research, to conclude on the clinical application of a model.

Additional information on validation strategies, seen from the perspective of regulatory science, can be found in Subheading 3 of Chap. 23. We focus here on an accessible discussion of the main concepts to have in mind concerning model evaluation strategies, and Raschka [8] gives a more mathematically detailed coverage of related topics.

3.1 Evaluating a Learning Procedure

We first focus on assessing the expected performance of a learning procedure on data drawn from a given population. Here, the model is validated on data with similar characteristics to the one used for training, a validation sometimes called *internal validation*. Most importantly, performance should not be evaluated using the same data that were used for training [6]. Therefore, the first step is to split the data into a training set and a testing set. This should be done before starting any work on the data, be it training a ML model or even doing simple statistics for identifying interesting features. Splitting the data can be done using `sklearn.model_selection.train_test_split` or `sklearn.model_selection.ShuffleSplit(n_splits=1)` from scikit-learn. When one simply performs a single split of the data into training and testing set, the validation method is called “hold-out.” One should nevertheless check that the training and testing sets have similar characteristics. More precisely, we want the output variable distribution to be approximately the same in the training and testing sets. This is called stratification. For instance, for classification, the proportion of diseased individuals should approximately be the same in the two sets. To that purpose, use `StratifiedShuffleSplit(n_splits=1)`. In medical applications, it is recommended to control not only for the disease status but also for other variables, such as sociodemographic information (age, sex, ...) or some relevant clinical variables. It will often be difficult (and it is not even necessary) to obtain almost identical distributions between training and testing sets. In practice, it is often

sufficient to have similar means and variances for continuous variables and similar proportions for categorical variables. The first two rows of Fig. 7 illustrate the concepts of “hold-out” and stratification.

Non-independent Samples. Prediction may be performed across non-independent data points, for instance, different points in a time series, or repeated measures of the same individual. In such case, it is important that samples in the train and test sets are independent, which may require selecting separated time windows. Also, the cross-validation should mimic the intended usage of the predictor. For instance, a diagnostic model intended to be applied to new individuals should be evaluated making sure that there are no shared individuals between the train and test sets.

3.1.1 Cross-validation

The split between train and test sets is arbitrary. With the same machine learning algorithm, two different data splits will lead to two different observed performances, both of which are noisy estimates of the expected generalization performance of prediction models built with this learning procedure. A common strategy to obtain better estimates consists in performing multiple splits of the whole dataset into training and testing sets: a so-called *cross-validation* loop. For each split, a model is trained using the training set, and the performances are computed using the testing set. The performances over all the testing sets are then aggregated. Figure 7 displays different cross-validation methods. k-fold cross-validation consists in splitting the data into k sets (called folds) of approximately equal size. It ensures that each sample in the dataset is used exactly once for testing. For classification, `sklearn.model_selection.StratifiedKFold` performs stratified k-fold cross-validation.

In each split, ideally, one would want to have a large training set, because it usually allows training better performing models, and a large testing set, because it allows a more accurate estimation of the performance. But the dataset size is not infinite. Splitting out 10–20% for the test set is a good trade-off [9], which amounts to $k = 5$ or 10 in a k-fold. With small datasets, to maximize the amount of train data, it may be tempting to leave out only one observation, in a so-called *leave-one-out* cross-validation. However, such depletion of the test set gives overall worse estimates of the generalization performance. Increasing the number of splits is, however, useful, and thus another strategy consists in performing a large number of random splits of the data, breaking from the regularity of the k-fold. If the number of splits is sufficiently large, all samples will be approximately used the same number of times for training and testing. This strategy can be done using `sklearn.model_selection.StratifiedShuffleSplit(n_splits)` and is called “Repeated hold-out” or “Monte-Carlo cross-validation.”

Whole data-set**Hold out**

Training set

Test set

Stratification

Training set

Test set

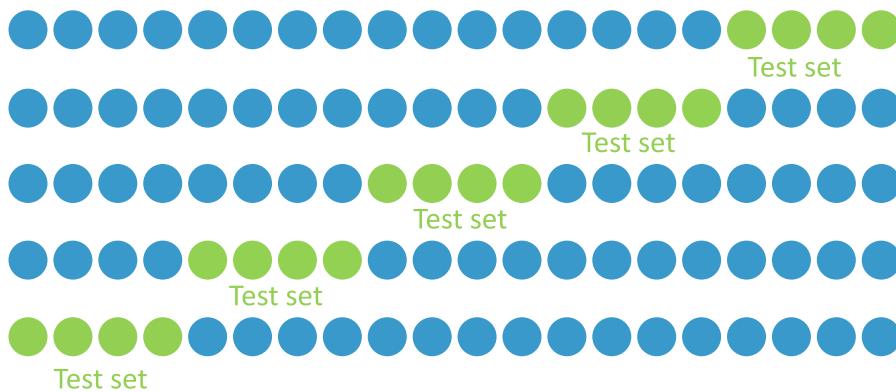
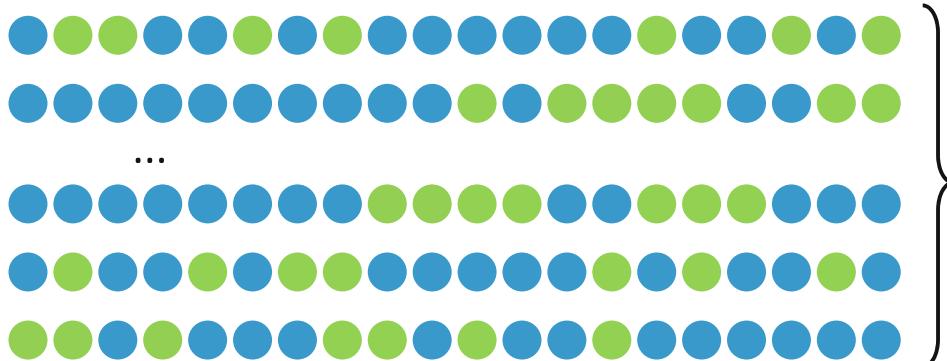
● Diseased ● Healthy
k-fold cross-validation (here k=5)**Repeated hold-out**
 Repeat a
large
number of
times

Fig. 7 Different validation methods, from top to bottom. The first method, called “hold-out,” involves a single split of the dataset into training and testing sets. It is thus not a cross-validation method. Stratification is the procedure that controls that the output variable (for instance, disease vs. healthy) has approximately the same distribution in the training and testing sets. k-fold cross-validation consists in splitting the data into k sets (called folds) of approximately equal size. Repeated hold-out consists in performing a large number of random splits of the data

Beyond giving a good estimate of the generalization performance, an important benefit of this strategy is that it enables to study the variability of the performances. However, running many splits may be computationally expensive with models that are slow to train.

3.1.2 The Need of an Additional Validation Set

Often, it is useful to make choices on the model to maximize prediction performance: make changes on the architecture, tune hyper-parameters, perform early stopping,.... As the test set performance is our best estimate of prediction performance, it would be natural to run cross-validation and pick the best model. However, in such a situation, the performances reported on the testing set will have an optimistic bias: a data-dependent choice has been made on this test set. There are two main solutions to this issue. The first one is usually applied when the model training is fast and the dataset is of small size. It is called nested cross-validation. It consists in running two loops of cross-validation, one nested into the other. The inner loop serves for hyper-parameter tuning or model selection, while the outer loop is used to evaluate the performance. The second solution is to separate from the whole dataset the test set, which will only be used to evaluate the performances. Then, the remainder of the dataset can be further split into training data and data used to make modeling choices, called the validation set.¹ Such a procedure is illustrated in Fig. 8. Commonly, the training and validation sets will be used in a cross-validation manner. They can then be used to experiment with different models, tune parameters,.... It is absolutely crucial that the test set is isolated at the very beginning, before any experiment is done. It should be left untouched and used only at the end of the study to report the performances. As for the split between training and validation sets, it is desirable that stratification is done when isolating the test set.

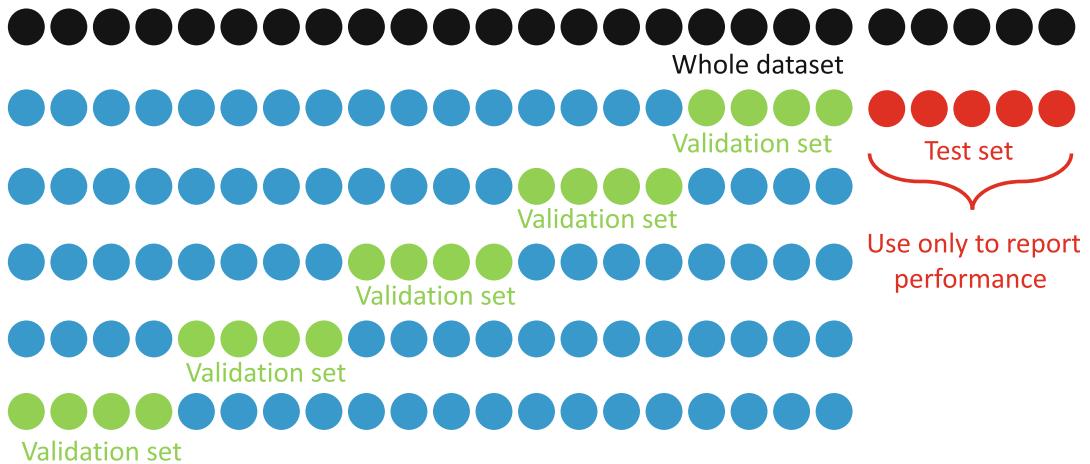
If the dataset is very small, nested cross-validation should be preferred as it gives better testing power than hold-out: all the data are used alternatively for model testing. If the dataset feels too small to split into train, validation, test, it may be too small to conduct a trustworthy machine learning study [10].

3.1.3 Various Sources of Data Leakage

Data leakage denotes cases where some information from the training set has “leaked” into the test set. As a consequence, the estimation of the performances is likely to be optimistic. Data leakage can be introduced in many ways, some of which are particularly insidious:

¹ In Chapter 23, the validation set is called the tuning set, as it is the standard practice in regulatory science and because it insists on the fact that it should not be used to evaluate the final performance, which should be done on an independent test set. In the present chapter, we use the term validation set as it is the most common in the academic setting.

Training, validation and test set



Use to train the model, experiment with different architectures...

Fig. 8 A standard approach consists in splitting the whole dataset into training, validation, and test sets. The test set must be isolated from the very beginning, left untouched until the end of the study and only be used to evaluate the performance. The training and validation sets are often used in a cross-validation manner. They can be used to experiment with different architectures and tune parameters

ous and may not be obvious to a researcher that is not familiar with a specific application field. Below, we describe some common causes of data leakage. A summary can be found in Box 3.

Box 3: Some Common Causes of Data Leakage

- Perform feature selection using the whole dataset.
- Perform dimensionality reduction using the whole dataset.
- Perform parameter selection using the whole dataset or the test set.
- Perform model or architecture search using the whole dataset or the test set.
- Report the performance obtained on the validation set that was used to decide when to stop training (in deep learning).
- For a given patient, put some of its visits in the training set and some in the validation set.
- For a given 3D medical image, put some 2D slices in the training set and some in the validation set.

A first basic cause of data leakage is to use the whole dataset for performing various operations on the data. A very common example is to perform feature selection using the whole dataset and then to use the selected features for model training. A similar situation is when dimensionality reduction is performed on the whole dataset. If this is done in an unsupervised manner (for example, using principal component analysis), it is likely to introduce less bias in the performance estimation because the target is not used. It nevertheless remains, in principle, a bad practice. A common practice in deep learning is to perform early stopping, i.e., use the validation set to determine when to stop training. If this is the case, the validation performances can be overoptimistic, and a separate test dataset should be used to report performance. Another cause of data leakage is when there are multiple longitudinal visits (i.e., the patient is evaluated at several time points) or multiple modalities for a given patient. In such a case, one should never put data from the same patient in both the training and validation sets. For instance, one should not, for a given patient, put the visit at month 0 in the training set and the visit at month 6 in the validation set. Similarly, one should not use the magnetic resonance imaging (MRI) data of a given patient for training and the positron emission tomography (PET) image for validation. A similar situation arises when dealing with 3D medical image. It is absolutely mandatory to avoid putting some of the 2D slices of a given patient in the training set and the rest of the slices in the validation set. More generally, in medical applications, the split between training and test sets should always be done at the patient level. Unfortunately, data leakage is still prevalent in many machine learning studies on brain disorders. For instance, a literature review identified that up to 40% of the studies on convolutional neural networks for automatic classification of Alzheimer's disease from T1-weighted MRI potentially suffered from data leakage [11].

3.1.4 Statistical Testing

Sources of Variance

Train–test splits, cross-validation, and the like seek to estimate the expected generalization performance of a learning procedure. Keeping test data rigorously independent from algorithm development minimizes the bias of this estimation. However, there are multiple sources of arbitrary variations in these estimates. The most obvious one is the intrinsic randomness of certain aspects of learning procedures, such as the random initial weights in deep learning. Indeed, while fixing the seed of the random number generator may remove the randomness on given train data, this stability is misleading given this choice is arbitrary and not representative of the overall behavior of the machine learning algorithm on the data distribution of interest [12]. A systematic study of machine learning benchmarks [13] shows that their most important sources of variance are:

Choice of test data/split.

A given test set is an arbitrary sample of the actual population that we are trying to generalize to. As a result, the corresponding

measure of performance is an imperfect estimate of the actual expected performance. Subheading 3.2, below, gives the resulting confidence intervals for a fixed test set. Using multiple splits, and thus multiple test sets, improves the estimation [13], though it makes computing confidence intervals hard [14].

Hyper-parameter optimization.

The choice of hyper-parameters is imperfect, for instance, because of limited resources to tune these hyper-parameters. Another attempt to tune hyper-parameter would lead to a slightly different choice. Thus benchmarks do not give an absolute characterization of a learning procedure but are muddled by imperfect hyper-parameters.

Random seeds.

As mentioned above, random choices in a learning procedure—initial weights, random drop-out for neural networks, or bootstraps in bagging—lead to uncontrolled fluctuations in benchmarking results that do not characterize the procedure’s ability to generalize to new data.

Conclusions Must Account for Benchmarking Variance

With all these sources of arbitrary variance, the question is: given benchmarks of a learning procedure performance, or improvement, is it likely to generalize reliably to new data or rather to be due to benchmarking fluctuations? Considering, for instance, the performance metrics in Table 1, it seems a safe bet to say that the convolutional neural network outperforms the two others but what about the difference between the two other models? From an application perspective, the question is whether this observed difference is likely to generalize to new data.

To answer this question, we must account for estimation error for the expected generalization performance from the different sources of uncontrolled variance in the benchmarks, as listed above. The first source of error comes from the limited sample size to test the predictions of the different learning procedures. Indeed, suppose that the testing set was composed of 100 samples. In that case, if only 3 more samples had been misclassified by the

Table 1

Accuracies obtained by different ML models on a binary classification task. Which model performs best? While it is quite likely that the convolutional neural network outperforms the two other models, it is less clear for the two other models. It seems that the support vector machine results in a slightly higher accuracy but is it due to random fluctuations in the benchmarks? Will the difference carry over to new data?

Model	Accuracy
Logistic regression	0.72
Support vector machine	0.75
Convolutional neural network	0.95

support vector machine, the two models would have had the same performance. A difference of 3 out of 100 could be easily due to having drawn 3 samples not representative of the population. Other sources of variance are due to how stable the learning pipeline is: sensitivity to hyper-parameters, random initialization, etc.

Box 4: Statistical Procedure to Characterize a Learner

1. Perform k runs of:
 - (a) Randomly splitting out a test set
 - (b) Training the learning procedure on the train set
 - (c) Measuring the performance p on the test set

Choose different values of arbitrary parameters (such as random seeds) on each run, and if enough computing power, run hyper-parameter optimization each time. This results in a set of performance measures $\mathcal{M} = \{m_1, \dots, m_k\}$.
2. Use all the values $\{m_1, \dots, m_k\}$ to conclude on the performance of the learner:

Confidence intervals are given by percentiles of \mathcal{M} .
Standard deviation of \mathcal{M} can be used to gauge typical variance of performance, as it requires performing a smaller number of runs k than percentiles. Standard error should not be used (see text).

Learner comparison can be done by comparing two such set of values \mathcal{M} and \mathcal{M}' , typically counting the fractions of values in \mathcal{M} that outperform \mathcal{M}' (without any pairing). Statistical procedures such as t-test should not be used (see text).

A Simple Statistical Testing Procedure

Training and testing a prediction pipeline multiple times are needed to estimate the variability of the performance measure. The simplest solution is to do this several times while varying the arbitrary factors, such as split between the train and the test or random initialization (*see Box 4*). The resulting set of performance measures is similar to bootstrap samples and can be used to draw conclusions on the distribution of performances in a test set. Confidence intervals can be computed using percentiles of this distribution. Two learning procedures can be compared by counting the number of times that one outperforms the other: outperforming 75% of the times is typically considered as a reliable improvement [13]. If the available computing power enables training learning procedures only a few times, empirical standard deviations should be used, as they require less runs to estimate. The improvements brought by a learning procedure can then be compared to these standard deviations.

Note these procedures do not perform classic null-hypothesis significance testing, which is difficult here. In particular, the standard error across the various runs should not be used instead of the standard deviation: the standard error is the standard deviation divided by the number of runs. The number of runs can be made arbitrarily large given enough compute power, thus making the standard error arbitrarily small. But in no way does the uncertainty due to the limited test data vanish. This uncertainty can be quantified for a fixed test set—*see Subheading 3.2*, but in repeated splits or cross-validation, it is difficult to derive confidence intervals because the runs are not independent [14, 15]. In particular, *it is invalid to use a standard hypothesis test—such as a T-test—across the different folds of a cross-validation*. There are some valid options to perform hypothesis testing in a cross-validation setting [14, 16], but they must be implemented with care.

Another reason not to rely on null-hypothesis testing is that their statistical significance only asserts that the expected performance—or improvement—is non-zero over a test population of infinite size. From a practical perspective, we care about meaningful improvements on test sets of finite size, which is related to the notion of *acceptance* tests—as opposed to *significance*—in the Neyman–Pearson framework of statistical testing [17]. Unlike null-hypothesis significance testing, it requires choosing a non-zero difference considered as acceptable, for instance as implicitly set by considering that a new learning procedure should improve upon an existing one 75% of the times—far from chance, which lies at 50%.

3.2 Generalization to an External Population

The Importance of External Validation

The procedures described above characterize the expected error of a learning procedure applied on a given population. A related, but different, question is that of characterizing the error

of a given predictive model, typically output by a training machine learning procedure on a study population. That second question, related to the notion of *external validity*, is important for two reasons. First, it characterizes the specific predictive model that will be used in practice, “in production.” Indeed, variance in the learning procedure will lead to arbitrary variation in model performance as large as typical improvements achieved by developing better models [13]. Second, characterizing the model on the *target* population may be important, as it may differ markedly from the study population. Indeed, the techniques in the previous section rely on splitting the initial dataset in training and testing (or validation) sets; hence, these different sets are by construction drawn from the same population and have similar characteristics (data coming from the same hospital/centers/countries, similar age/sex,...). They only demonstrate the ability of the model to generalize to new but similar data. To better assess model utility, guidelines on evaluating clinical prediction models insist on external validation using data collected later in time, or in a different geographical area [18].

Testing whether a prediction model can generalize to dissimilar data is important as it is all too frequent that the study sample, on which the model was developed, does not represent the target population [19]. The target data may, for instance, come from different hospitals and different countries, be acquired with different acquisition devices and protocols or with different sociodemographic or clinical characteristics than those of the training data. For instance, it has been shown that the type of MRI scanner can have a substantial impact on the generalization ability of ML models. To assess such generalization ability, a common practice is to use one or several additional datasets for testing, these datasets being acquired using different protocols and at different sites (Fig. 9). Most often, these datasets come from other research studies (different from the one used for training). However, research studies do not usually reflect well clinical routine data. Indeed, in research studies, the acquisition protocols are often standardized and rigorous data quality control is applied. Moreover, participants may not be representative of the target population. This can be due to inclusion/exclusion criteria (for instance, excluding patients with vascular abnormalities in a study on Alzheimer’s disease) or due to uncontrolled biases. For instance, participants to research studies tend to have a higher socioeconomic status than the general population. Therefore, it is highly valuable to also perform validation on clinical routine data, whenever possible, as it is more likely to reflect “real-life” situations. One should nevertheless be aware that a given clinical routine dataset may come with specificities that may not generalize to all settings. For instance, data collected within a specialized center of a university hospital may substantially differ from that seen by a general practitioner.

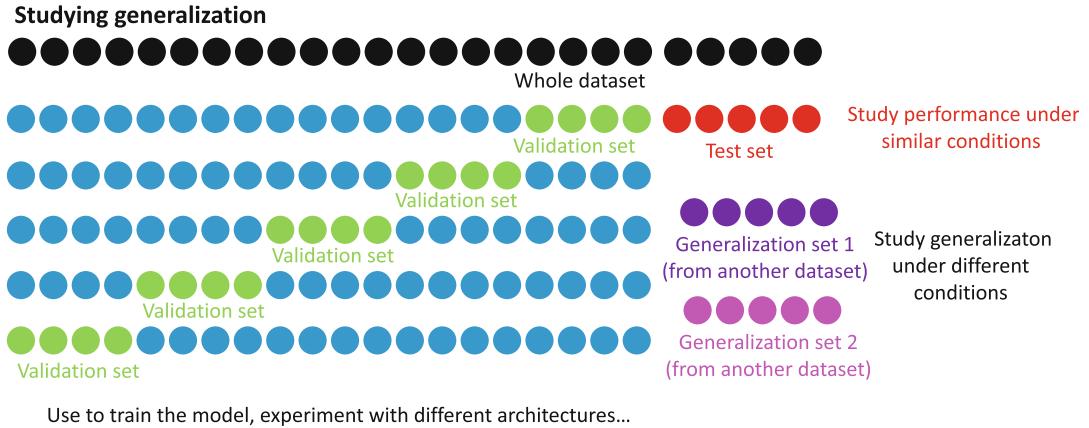


Fig. 9 In order to assess the generalization ability of a model under different conditions (such as data coming from different hospitals/countries, acquired with different devices and protocols...), a common practice is to use one or several additional datasets that come from other studies than the one used for training

Testing Procedures for External Validation

External validation of a predictive model relies on an independent test set and not cross-validation. Statistical testing thus amounts to derive confidence intervals or null-hypothesis significance testing for the metric of interest on this test set, exactly as when characterizing a diagnostic test [20].

For simple metrics that rely on counting successes, such as accuracy, sensitivity, PPV, NPV, the sampling distribution can be deduced from a binomial law. Table 2 gives such confidence intervals for a different set of the test set and different values of the ground-truth accuracy. These can be easily adapted to other counts of errors as follows:

Accuracy	N is the size of the test set
Sensitivity	N is the number of negative samples in the test set
Specificity	N is the number of positive samples in the test set
PPV	N is the number of positively classified test samples
NPV	N is the number of negatively classified test samples

We believe it is very important to have in mind the typical orders of magnitude reported in Table 2. It is not uncommon to find medical classification studies where the test set size is about a hundred or less. In such a situation, the uncertainty on the estimation of the performance is very high.

These parametric confidence intervals are easy to compute and refer to. But actual confidence intervals may be wider if the samples are not i.i.d. In addition, some interesting metrics, such as AUC ROC, do not come with such parametric confidence interval. A

Table 2
Binomial confidence intervals on accuracy (95% CI) for different values of ground-truth accuracy

N	65%	80%	90%	95%
100	[-9.0% 9.0%]	[-8.0% 8.0%]	[-6.0% 5.0%]	[-5.0% 4.0%]
1000	[-3.0% 2.9%]	[-2.5% 2.4%]	[-1.9% 1.8%]	[-1.4% 1.3%]
10,000	[-0.9% 0.9%]	[-0.8% 0.8%]	[-0.6% 0.6%]	[-0.4% 0.4%]
100,000	[-0.3% 0.3%]	[-0.2% 0.2%]	[-0.2% 0.2%]	[-0.1% 0.1%]

general and good option, applicable to all situations, is to approximate the sampling distribution of the metric of interest by bootstrapping the test set [8].

Finally, note that all these confidence intervals assume that the available labels are the ground truth. In practice, medical truth is difficult to establish, and label error may bias the estimation of error rates.

When comparing two classifiers, a McNemar’s test is useful to test whether the observed difference in errors can be explained solely by sampling noise [21, 22]. The test is based on the number of samples misclassified by one classifier and not the other, n_{01} and vice versa n_{10} . The test statistics is then written $(|n_{01} - n_{10}| - 1)^2 / (n_{01} + n_{10})$; it is distributed under the null as a χ^2 with 1 degree of freedom. To compare classifiers scanning the trade-off between specificity and sensitivity without choosing a specific threshold on their score, one option is to compare areas under the curve of the ROC, using the DeLong test [23] or a permutation scheme to define the null [24].

4 Conclusion

Evaluating machine learning models is crucial. Can we claim that a new model outperforms an existing one? Is a given model trustworthy enough to be “deployed,” making decisions in actual clinical settings? A good answer to these questions requires model evaluation experiments adapted to the application settings. There is no one-size-fits-all solution. Multiple performance metrics are often important, chosen to reflect target population and cost-benefit trade-offs of decisions, as discussed in Subheading 2. The prediction model must *always* be evaluated on unseen “test” data, but different evaluation goals lead to procedures to choose these test data. Evaluating a “learner”—a model construction algorithm—leads to cross-validation, while evaluating the fitness

of a given prediction rule—as output by model fitting—calls for left-out data representative of the target population. In all settings, accounting for uncertainty or variance of the performance estimate is important, for instance, to avoid investing in models that bring no reliable improvements.

Acknowledgements

This work was supported by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6), ANR-20-CHIA-0026 (LearnI). We thank Sebastian Raschka for detailed feedback.

Appendix

Odds Ratio and Diagnostic Tests Evaluation

Odds

Odds and odds ratio are frequently used in biostatistics and epidemiology, but less in machine learning. Here we give a quick introduction to these topics.

Odds are a measure of likelihood of an outcome: the ratio of the number of events that produce that outcome to the number that do not. The odds $\mathcal{O}(\alpha)$ of an outcome α are simply related to the probability $P(\alpha)$ of this outcome:

$$\text{Odds of } \alpha \quad \mathcal{O}(\alpha) = \frac{P(\alpha)}{1 - P(\alpha)}. \quad (1)$$

In other words, $\mathcal{O}(\alpha)$ is the number of times the event α would occur for each occurrence of the opposite event. This intuitive explanation has led odds to be often used for sports gambling. For instance, if the odds are 3 (or more specifically in gambling terminology 3 : 1) for FC Barcelona vs. Real Madrid, it means that FC Barcelona has a probability of winning against Real Madrid of 75% ($P(\alpha) = \frac{\mathcal{O}(\alpha)}{\mathcal{O}(\alpha)+1}$). Coming back to diseases, supposing that only a minority of the population is affected, if the odds of the disease are 1%, which can be written as 1 : 100, this means that for every diseased person in the population, there are 100 persons without it. The prevalence is thus $\frac{1}{101} = 0.99\% \approx 1\%$. One can see that when the prevalence is low, it is close to the odds, which is not the case when prevalence gets higher. This is true in general of probabilities and odds: when the probability is low, it is close to the odds.

Odds Ratio and Invariance to Class Sampling

The odds ratio measures the association between two events, a and b , which we can arbitrarily call respectively outcome and property. The odds ratio is defined as the ratio of the odds of the outcome in the group where the property holds to that in the group where the property does not hold:

$$\text{Odds ratio between } a \text{ and } b \quad OR(a, b) = \frac{\mathcal{O}(a|b=+)}{\mathcal{O}(a|b=-)}. \quad (2)$$

To compute the odds ratio, the problem is fully specified by the counts in the following contingency table:

		Outcome a	
		$a+$	$a-$
Property b	$b+$	n_{++}	n_{-+}
	$b-$	n_{+-}	n_{--}

(3)

The odds are written: $\mathcal{O}(a|b=+) = \frac{n_{++}}{n_{-+}}$ and $\mathcal{O}(a|b=-) = \frac{n_{+-}}{n_{--}}$; hence, the odds ratio reads

$$OR(a, b) = \frac{n_{++}}{n_{-+}} \frac{n_{--}}{n_{+-}}. \quad (4)$$

Note that this expression is unchanged swapping the role of a and b ; the odds ratio is symmetric, $OR(a, b) = OR(b, a)$.

Invariance to Class Sampling

Suppose we have sampled the population selecting with a frequency f on the outcome $a+$, for instance, to oversample the positive outcome or the positive property.² In Eq. 4, n_{++} is replaced by $f n_{++}$ and n_{+-} by $f n_{+-}$; however, the factor f cancels out and the overall expression of the odds ratio is unchanged. This is a central property of the odds ratio:

The odds ratio is unchanged by sample selection bias on one of the variables (a or b).

This property is one reason why odds and odds ratio are so central to biostatistics and epidemiology: sampling or recruitment bias is an important concern in these fields. For instance, a case-control study has a very different prevalence as the target population, where the frequency of the disease is typically very low.

²Indeed, thankfully, many diseases have a prevalence much lower than 50%, e.g., 1%, which is already considered a frequent disease. Therefore, in order to have a sufficient number of diseased individuals in the sample without dramatically increasing the cost of the study, diseased participants will be oversampled. One extreme example, but very common in medical research, is a case-control study where the number of diseased and healthy individuals is equal.

Confusion with Risk Ratio

The odds ratio is often wrongly interpreted as a risk ratio—or relative risk—which is more easily understood.

The risk ratio is the ratio of the probability of an outcome in a group where the property holds to the probability of this outcome in a group where this property does not hold. The risk ratio thus differs from the odds ratio in that it is expressed for probabilities and not odds. Even though the values for odds ratio and risk ratio are often close because, in most diseases being diseased is much less likely than not, they are fundamentally different because the odds ratio does not depend on sampling whereas the risk ratio does.

Likelihood Ratio of Diagnostic Tests or Classifiers

The likelihood ratio used to characterize diagnostic tests or classifiers is strongly related to the odds ratio introduced above, though it is not strictly speaking an odds ratio. It is defined as

$$\text{LR}+ = \frac{P(T+|D+)}{P(T-|D-)}. \quad (5)$$

Using the expressions in Box 1 and the fact that $P(T-|D-) = 1 - P(T+|D-)$, the $\text{LR}+$ can be written as

$$\text{LR}+ = \frac{\text{Sensitivity}}{1 - \text{Specificity}}. \quad (6)$$

Link to Pre-test and Post-test Odds

We can write this in terms of the contingency table in Eq. 3 (the link to the confusion matrix in Fig. 1 is given by $a = D$, $b = T$ and thus $n_{++} = TP$, $n_{+-} = FP$, $n_{-+} = FN$, $n_{--} = TN$):

$$\text{LR}+ = \frac{n_{++}}{n_{++} + n_{+-}} \frac{n_{-+} + n_{--}}{n_{-+}} \quad (7)$$

$$= \underbrace{\frac{n_{++}}{n_{-+}}}_{\frac{P(D+|T+)}{P(D-|T+)}} \underbrace{\frac{n_{-+} + n_{--}}{n_{++} + n_{-+}}}_{\frac{P(D-)}{P(D+)}} \quad (8)$$

$\frac{P(D+|T+)}{P(D-|T+)} = O(D+|T+)$ $\frac{P(D-)}{P(D+)} = \frac{1}{O(D+)}$

$$\text{LR}+ = \frac{O(D+|T+)}{O(D)}. \quad (9)$$

Indeed, $O(D+|T+) = \frac{P(D+|T+)}{1 - P(D+|T+)} = \frac{P(D+|T+)}{P(D-|T+)}$ and $O(D) = \frac{P(D)}{1 - P(D)} = \frac{P(D)}{P(D-)}$.

$O(D+)$ is called the pre-test odds (the odds of having the disease in the absence of test information). $O(D+|T+)$ is called the post-test odds (the odds of having the disease once the test result is known).

Equation 9 shows how the LR+ relates pre- and post-test odds, an important aspect of its practical interpretation.

Invariance to Prevalence

If the prevalence of the population changes, the quantities are changed as follows: $n_{++} \rightarrow f n_{++}$, $n_{+-} \rightarrow f n_{+-}$, $n_{-+} \rightarrow (1-f) n_{-+}$, $n_{--} \rightarrow (1-f) n_{--}$, affecting LR+ as follows:

$$\text{LR}+ = \frac{f n_{++}}{(1-f) n_{-+}} \frac{(1-f) n_{-+} + (1-f) n_{--}}{f n_{++} + f n_{+-}}. \quad (10)$$

The factors f and $(1-f)$ cancel out, and thus the expression of LR+ is unchanged for a change of the pre-test frequency of the label (prevalence of the test population). This is alike odds ratio, though the likelihood ratio is not an odds ratio (and does not share all properties; for instance, it is not symmetric).

References

1. Pedregosa F, et al (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12(85): 2825–2830
2. Powers D (2011) Evaluation: from precision, recall and f-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol* 2(1):37–63
3. Naeini MP, Cooper G, Hauskrecht M (2015) Obtaining well calibrated probabilities using Bayesian binning. In: Twenty-Ninth AAAI Conference on Artificial Intelligence
4. Vickers AJ, Van Calster B, Steyerberg EW (2016) Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 352:i6
5. Perez-Lebel A, Morvan ML, Varoquaux G (2023) Beyond calibration: estimating the grouping loss of modern neural networks. In: ICLR 2023 Conference
6. Poldrack RA, Huckins G, Varoquaux G (2020) Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* 77(5): 534–540
7. Barocas S, Hardt M, Narayanan A (2019) Fairness and machine learning. <http://www.fairmlbook.org>
8. Raschka S (2018) Model evaluation, model selection, and algorithm selection in machine learning. Preprint arXiv:181112808
9. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B (2017) Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage* 145:166–179
10. Varoquaux G (2018) Cross-validation failure: small sample sizes lead to large error bars. *NeuroImage* 180:68–77
11. Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Dormont D, Durrleman S, Burgos N, Colliot O, et al (2020) Convolutional neural networks for classification of Alzheimer’s disease: overview and reproducible evaluation. *Med Image Anal* 63:101694
12. Bouthillier X, Laurent C, Vincent P (2019) Unreproducible research is reproducible. In: International Conference on Machine Learning, PMLR, pp 725–734
13. Bouthillier X, Delaunay P, Bronzi M, Trofimov A, Nichyporuk B, Szeto J, Mohammadi Sepahvand N, Raff E, Madan K, Voleti V, et al (2021) Accounting for variance in machine learning benchmarks. *Proc Mach Learn Syst* 3:747–769
14. Bates S, Hastie T, Tibshirani R (2021) Cross-validation: what does it estimate and how well does it do it? Preprint arXiv:210400673
15. Bengio Y, Grandvalet Y (2004) No unbiased estimator of the variance of k-fold cross-validation. *J Mach Learn Res* 5(Sep):1089–1105
16. Nadeau C, Bengio Y (2003) Inference for the generalization error. *Mach Learn* 52(3): 239–281

17. Perezgonzalez JD (2015) Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Front Psychol* 6:223
18. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration. *Ann Int Med* 162(1):W1–W73
19. Dockès J, Varoquaux G, Poline JB (2021) Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience* 10(9): giab055
20. Shapiro DE (1999) The interpretation of diagnostic tests. *Statist Methods Med Res* 8(2): 113–134
21. Leisenring W, Pepe MS, Longton G (1997) A marginal regression modelling framework for evaluating medical diagnostic tests. *Statist Med* 16(11):1263–1281
22. Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput* 10(7): 1895–1923
23. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44:837–845
24. Bandos AI, Rockette HE, Gur D (2005) A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. *Statist Med* 24(18):2873–2893

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 21

Reproducibility in Machine Learning for Medical Imaging

Olivier Colliot, Elina Thibeau-Sutre, and Ninon Burgos

Abstract

Reproducibility is a cornerstone of science, as the replication of findings is the process through which they become knowledge. It is widely considered that many fields of science are undergoing a reproducibility crisis. This has led to the publications of various guidelines in order to improve research reproducibility.

This didactic chapter intends at being an introduction to reproducibility for researchers in the field of machine learning for medical imaging. We first distinguish between different types of reproducibility. For each of them, we aim at defining it, at describing the requirements to achieve it, and at discussing its utility. The chapter ends with a discussion on the benefits of reproducibility and with a plea for a nondogmatic approach to this concept and its implementation in research practice.

Key words Reproducibility, Replicability, Reliability, Repeatability, Open science, Machine learning, Artificial intelligence, Deep learning, Medical imaging

1 Introduction

Reproducibility is at the core of the scientific method. In its general and most common meaning, it corresponds to the ability to reproduce the findings of a given experimental study. This is a necessary (but not sufficient) condition for a scientific statement to become accepted as new knowledge. Let's illustrate this with a simple example, considering the following statement: “the volume of the hippocampus is, on average, smaller in patients with Alzheimer's disease (AD) than in healthy people of comparable age.” Such statement was the conclusion of studies which measured such volume from magnetic resonance images (MRI). To the best of our knowledge, the first study to assert this was that of Seab et al [1]. This was later reproduced by many other studies (e.g., [2, 3]). It is now widely accepted, which would not have been the case if the study had proven impossible to reproduce. Note that, as stated above, this is a *necessary* but not a *sufficient* condition. Indeed, there could be other reasons for such statement not to be considered as knowledge. For instance, let's imagine that some other

researchers discover that there is an artifact that is systematically present in the MRI of patients with AD and which leads to erroneous volume estimation. Then, the statement could not be considered new knowledge even though it had been reproduced several times.

Machine learning (ML) is, in part, an experimental science. This is not the case of the entirety of the discipline, part of which is theoretical (for instance, mathematical proofs of convergence or of approximation capabilities of different classes of models) or methodological (the invention of a new approach). Nevertheless, since ML ultimately aims at solving practical problems, its experimental component is essential. Typically, one would want to be able to make statements of the type described above from an experimental study. Here is an example of such statement: “this ML model (for instance, a specific convolutional neural network [CNN] architecture), using MRI data as input, is capable of classifying AD patients and healthy controls with an accuracy superior to 80%.” In order to end an article with such a statement, one needs to conduct an experimental study. For such findings to become knowledge, it needs to be subsequently reproduced. Of course, this statement is unlikely to be universal, and one would want to know under which conditions it holds: for instance, is it restricted to a specific class of MRI scanners, to specific disease stages, to specific age ranges?

Box 1: Glossary

The readers will find the definition of the terms we used in the present document.

- **Reproducibility, replicability, repeatability.** In the present document, these will be used as synonyms of reproducibility.
- **Original study.** Study that first showed a finding.
- **Replication study.** Study that subsequently aimed at replicating an original study, with the hope to support its findings.
- **Research artifact.** Any output of scientific research: papers, code, data, protocols.... Not to be confused with imaging artifacts which are defects of imaging data.
- **Claims.** The conclusions of a study. Basically a set of statements describing the results and a set of limitations which delineate the boundaries within which the claims are stated (the term “claim” is here used in the broad scientific sense not with the specific meaning it has in the context of regulation of medical devices although the two may be related).
- **Limitations.** A set of restrictions under which the claims may not hold (usually because the corresponding settings have not been explored).

(continued)

Box 1 (continued)

- **Method.** The ML approach described in the paper, independently of its implementation.
- **Code.** The implementation of the method.
- **Software dependencies.** Other software packages that the main code relies on and which are necessary for its execution.
- **Public data.** Data that can be accessed by anybody with no or little restriction (for instance, the data hosted at <https://openneuro.org>).
- **Semi-public data.** Data which requires approval of a research project (for instance, the Alzheimer’s Disease Neuroimaging Initiative [ADNI] <http://www.adni-info.org>). The researchers can then use the data only for the intended research purpose and cannot redistribute it.
- **Data split.** Separation into training, validation, and test sets.
- **Data leakage.** Faulty procedure which has led information from the training set to leak into the test set. *See refs. 4, 5 for details.*
- **Error margins.** A general term for providing the precision of the performance estimates (e.g., standard-error or confidence intervals).
- **Researcher degrees of freedom.** Number of different components (e.g., different architectures, hyperparameter values, subsamples...) which have been tried before arriving to the final method [6]. Too many degrees of freedom tend to produce methods that do not generalize.
- **p-hacking.** A bad practice that involves too many degrees of freedom and which consists is trying many different statistical procedures until a significant p-value is found.
- **Acquisition settings.** Factors that influence the scan of a given patient (imaging device, acquisition parameters, image quality).
- **Image artifacts.** Defects of a medical image, these may include noise, field heterogeneity, motion artifacts, and others.
- **Preregistration.** The deposit of the study protocol prior to performing the study. Limits degrees of freedom and increases likelihood of robust findings.

In the examples above, we have actually illustrated only one on the many possible meanings of reproducibility: the addition of new evidence to support a scientific finding of an *original* study through reproduction under different experimental conditions (*see Box 1* for a glossary of some of the key concepts used). However, it is also used for very different meanings. In computational sciences, it is

often used for the ability to exactly reproduce the results (i.e., the exact numbers) in a given study. In sciences which aim at providing measurements (as is often the case in medical imaging), the word may be used to describe the variability of a given measurement tool under different acquisition settings. We shall provide more details on these different meanings in Subheading 2. Finally, the topics of reproducibility and open science are obviously related since the latter favors the former. However, open science encompasses a broader objective which is to make all *research artifacts* (code, data, papers...) openly available for the benefit of the whole society. Conversely, open research may still be unreproducible (e.g., because it has relied on faulty statistical procedures).

There has been increasing concern that science is undergoing a reproducibility crisis [7–10]. This is present in various fields from psychology [11] to preclinical oncology research [12]. ML [13–17], digital medicine [18], ML for healthcare [19, 20], and ML for medical imaging [21] are no exception. The concerns are multifaceted. In particular, they include two substantially different aspects: the report of failures to reproduce previous studies and the observation that many papers do not provide sufficient information for reproducing their results. It is important to have in mind that, while the two may be related, there is not a direct relationship between them: it may very well be that a paper seems to include all the necessary information for reproduction and that reproduction attempts fail (for instance, because the original study had too many degrees of freedom and led to a method that only works on a single dataset, *see* Subheading 4).

Various guidelines have been proposed to improve research reproducibility. Such guidelines may be general [10] or devoted to specific fields including brain imaging [22–25] and ML for healthcare and life sciences [26, 27]. Moreover many other papers, even though not strictly providing guidelines, provide very valuable pieces of advice for making research more trustworthy and in particular more reproducible (e.g., [14, 19, 28–32]).

This chapter is an introduction to the topic of reproducibility for researchers in the field of ML for medical imaging. It is not meant at providing a replacement for the aforementioned previously published guidelines that we strongly encourage the reader to refer to.

The remainder of the chapter is organized as follows. We first start by introducing different types of reproducibility (Subheading 2). For each of them, we attempt to clearly define it and describe what are the requirements to achieve it and the benefits it can provide (Subheadings 3, 4, 5, and 6). All this information is given with having the field of ML for medical imaging as a target, even though part of it may apply to other fields. Finally, we conclude with a discussion which both describes the benefits of reproducibility but also advocates for a nondogmatic point of view on the topic (Subheading 7).

2 The Polysemy of Reproducibility

The term “reproducibility” has been used with various meanings which may range from the exact reproduction of a study with the same material and methods, to the reproduction of a result using new experimental data to the support of a scientific idea using a completely different experimental setup [33, 34]. Moreover, various terms have been introduced including reproducibility, replicability, repeatability, reliability, robustness, generalizability... Some of these words, for instance, reproducibility vs replicability, have even been used by some authors with opposite meanings [33, 34]. We will not aim at assigning an unambiguous meaning to each of these words, as we find this of little interest, and will use the term “reproducibility,” “replicability,” and “repeatability” as synonyms. On the other hand, we believe, as many other authors [19, 23, 33, 35], that it is important to distinguish between different types of reproducibility. To that purpose, it is useful to have a *taxonomy* of reproducibility. Below, we describe such a taxonomy. We do not claim that it is novel, as it takes inspiration from other papers [14, 19, 23, 33, 35] nor that it should be universally adopted. Furthermore, boundaries between different types of reproducibility are partly fuzzy. We simply hope that it will be useful for the different concepts that we subsequently introduce and that it will be well adapted to the field of ML for medical imaging.

We distinguish between four main types of reproducibility: **exact reproducibility**, **statistical reproducibility**, **conceptual reproducibility**, and **measurement reproducibility**. We describe those four main types in the following sections. They are also summarized in Fig. 1. As will be explained below, the three first types have relationships with each other (this is why they have the same color in the figure) while the fourth is more separated.

3 Exact Reproducibility

What Is It? Exact reproducibility aims at reproducing strictly identical results as those of a previously published paper. Concretely, this amounts to being able to reproduce tables and figures as they appear in the original paper following the same procedures as the authors.

What Does It Require? Exact reproducibility requires to have access to all components that led to the results including, of course, data and code.

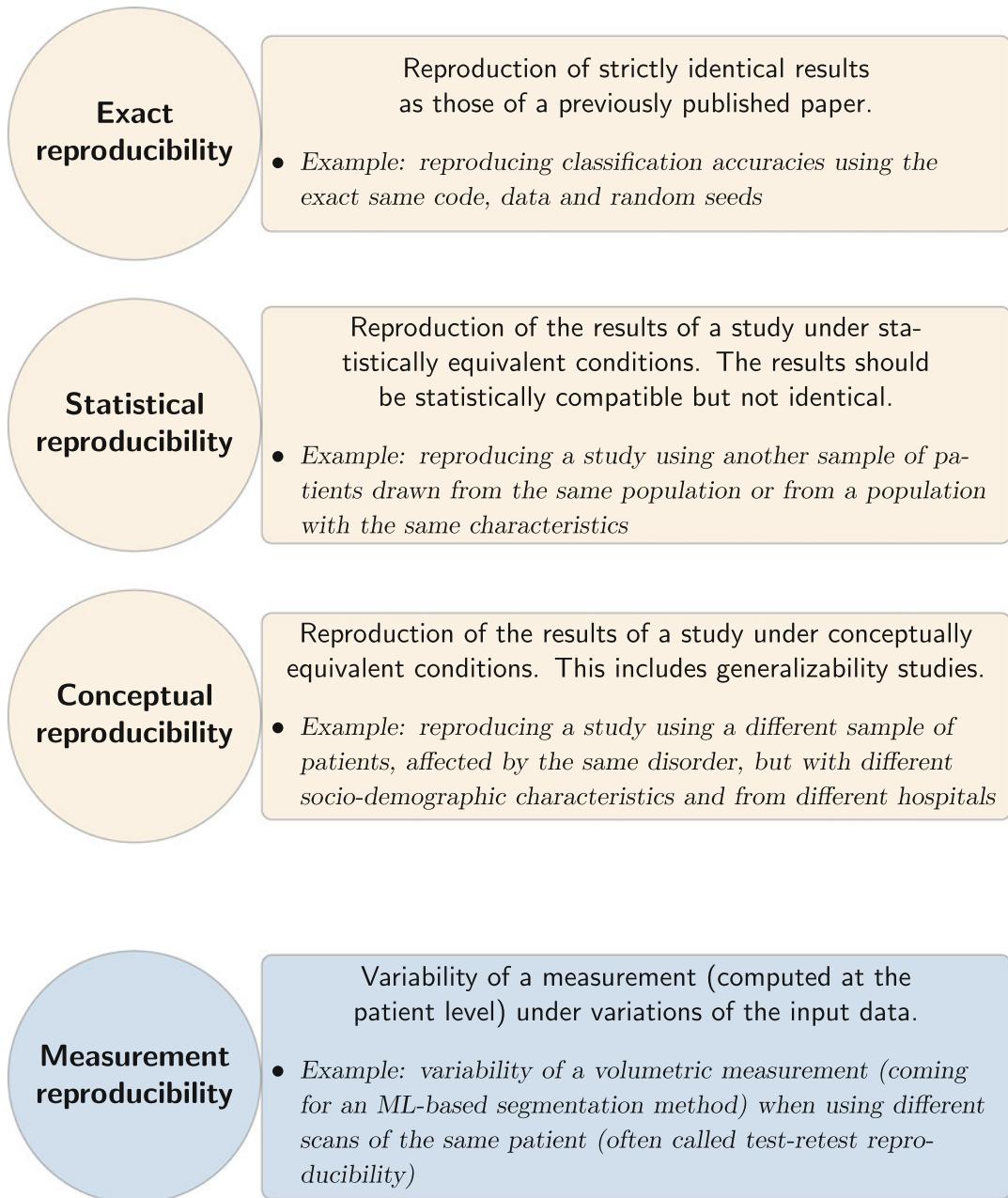


Fig. 1 Different types of reproducibility. Note that, in the case of “statistical” and “conceptual” reproducibility, the terms come from [19] but the exact definition provided in each corresponding section may differ

Access to data is obviously necessary [19, 22, 27]. Open data has been described (together with code and papers) as one of the pillars of open science [22]. It is widely accepted that scientific data should adhere to the FAIR (Findable, Accessible, Interoperable, Reusable) principles (please refer to <https://www.go-fair.org/fair-principles>

and [36] for more details). Among these principles, accessibility is often the most difficult to adhere to for medical imaging data (or healthcare data in general). It is very common in medical papers that data is mentioned as available upon request. However, a study has showed that, when data is subsequently requested, many researchers actually do not comply with the data accessibility statement [37]. This is worrisome, and more transparent ways of data sharing would be welcome. However, as mentioned above, such transparent sharing procedures may be difficult to put in place for healthcare data. In particular, making the data public is often difficult due to regulatory and privacy constraints [19]. Gorgolewski and Poldrack [22] provide useful pieces of advice to facilitate sharing, but there are cases where public sharing will remain impossible. In particular, one must distinguish between research data (acquired as part of a research protocol), which can often be made *public* or *semi-public*¹ provided that adequate measures have been taken at data collection (e.g., adequate participant consent), and routine clinical data (acquired as part of the routine clinical care of the patients), which sharing can be much more complicated. It is important that data is easily findable and that it is shared on a server which has a long-term maintenance. General purpose data repositories such as Zenodo² provide a good solution. Another important aspect is to adhere to community standards for data organization, so that it can easily be reused by researchers. For brain imaging, the community standard is BIDS (Brain Imaging Data Structure) [38].³ This standard is already very mature and has already been extended to incorporate other modalities such as microscopy images, for instance (Microscopy-BIDS [39]). Note that there is an ongoing proposal to extend it to other organs (MIDS – Medical Imaging Data Structure [40])⁴. Finally, we would like to draw the attention to an important point that is often overlooked. Even when a study relies on public or semi-public data, it is absolutely necessary to specify which samples (e.g., which participants and which scans) have been used; otherwise, the study is not reproducible [41]. Ideally, one would provide code to automatically make the data selection [42] in order to ease the replication.

Another key component is that the code is accessible [19, 22, 27]. Indeed, it would be delusional to think that exactly the same results could be obtained using a reimplementation based on information provided in the paper (even though it is good practice to provide as much detail as possible about the methods in the paper).

¹ See glossary Box 1.

² <https://zenodo.org/>.

³ <https://bids.neuroimaging.io/>.

⁴ See BIDS extension proposal (BEP) number 25 (BEP025) https://bids.neuroimaging.io/get_involved.html#extending-the-bids-specification

Theoretically, it does not mean that the code must come with an open software license. However, doing so has many additional benefits such as allowing other researchers to use the code or parts of it for different purposes. The code should be made according to good coding practices which include the use of a versioning system and adequate documentation [20]. Furthermore, although not strictly needed for reproducibility, the use of continuous integration makes the code more robust and eases its long-term maintenance. Besides, it is good to ease as much as possible the installation of dependencies [27]. This can be done with pip⁵ when programming in Python. One can also use containers such as Docker.⁶ One can find useful additional advice in the *Tips for Publishing Research Code*.⁷ Note that we are not saying that all these components should be present in any study or are prerequisites for good research. They constitute an ideal reproducibility goal.

Sharing well-curated notebooks is also a way to ease reproducibility of results by other researchers. This can be done through standard means, but dedicated servers also exist. One can, for instance, cite an interesting initiative called NeuroLibre⁸ which provides a preprint server for reproducible data analysis in neuroscience, in particular providing curated and reviewed Jupyter notebooks [43].

In ML, sharing the code itself is not enough for exact reproducibility. First, every element of the training procedure should be stored: this includes the data splits and the criteria for model selection. Moreover, there usually are non-deterministic components so it is necessary to store random seeds [27]. Furthermore, software/operating system versions, the GPU model/version, and threading have been deemed necessary to obtain exact reproducibility [44]. The ClinicaDL software platform provides a framework for easing exact reproducibility of deep learning for neuroimaging [5].⁹ Although it is targeted at brain imaging, many of its components and concepts are applicable to medical imaging in general. Also in the field of brain imaging, NiLearn¹⁰ facilitates the reproducibility of statistics and ML. One can also cite Pymia¹¹ which provides data handling and validation functionalities for deep learning in medical imaging [45].

⁵ <https://pypi.org/project/pip/>.

⁶ <https://www.docker.com/>.

⁷ <https://github.com/paperwithcode/releasing-research-code>.

⁸ <https://neuroibre.org/>.

⁹ <https://clinicadl.readthedocs.io/en/latest/>.

¹⁰ <https://nilearn.github.io/stable/index.html>.

¹¹ <https://github.com/rundherum/pymia>.

Finally, it may seem obvious, but, even when the code is shared, the underlying theory of the method, all its components, and implementation details need to be present in the paper [14].

It is in principle possible to retrain models identically if the above elements are provided. It nevertheless remains a good practice to share trained models, in order to allow other researchers to check that retraining indeed led to the same results but also to save computational resources. However, models can be attacked to recover training data [27, 46]. This is not a problem when the training data is public. When it is privacy-sensitive, methods to preserve privacy exist [27, 47].

In medical imaging, preprocessing and feature extraction are often critical steps that will subsequently influence the ML results. It is thus necessary to also provide code for such parts. Several software initiatives including BIDSApps [48]¹² and Clinica [49]¹³ provide ready-to-use tools for preprocessing and feature extraction for various brain imaging modalities. Applicable to many medical imaging modalities, the ITK [50, 51]¹⁴ framework provides a wide range of processing tools. It can ease the work of researchers who do not want to spend time on preprocessing and feature extraction pipelines and focus on the ML part of their work.

Why Is It Useful? It has been claimed that exact reproducibility is of little interest, that pursuing it is a waste of energy of the community, and that its only possible use would be the detection of outright fraud which is rare [52]. We disagree with that view. Let’s start with fraud. It may be of low occurrence, although its exact prevalence is difficult to establish. Even so, it is of disastrous consequences as it leads to loss of trust by students, scientists, and the general public. In particular, a survey of 1,576 researchers indicated that 40% of them believe that fraud is a factor that “always/often” contributes to irreproducible research and that 70% of them think that it “sometimes” contributes [7]. Exact reproducibility can certainly contribute to reduce fraud as full transparency obviously makes fraud more difficult. Fraud remains possible (one could forge some data and share it), but it is more difficult to achieve under transparency constraints. Fraud may be rare but errors are much more common. The framework of exact reproducibility eases the detection of errors which is a service to science and even to the authors themselves. In particular, it may help discover “biases and artifacts in the data that were missed by the authors and that cannot be discovered if the data are never made available” [27]. Similarly, it can lead to the discovery of

¹² <https://bids-apps.neuroimaging.io/>.

¹³ <https://www.clinica.run/>.

¹⁴ <https://itk.org/>.

wrong validation schemes, including data leakage or errors in implementation that make it inconsistent with the methodology presented in the paper. Overall, it may make progress slower, but it will definitely make it steadier. However, this does not mean that exact reproducibility should be aimed in all works or made a requirement for all publications (*see* Subheading 7.4).

4 Statistical Reproducibility

What Is It? Statistical reproducibility aims at reproducing findings under statistically equivalent conditions.¹⁵ The specific definition may vary, but the following choices are often considered reasonable. The implementation of the method (the code) remains the same. Random components are left random. Regarding the data, the general idea is that the sample would be drawn from the same population. One could, for instance, use subsamples of the original data or another subsample of a larger source population. An interesting case is to study different data splits. A less restrictive view of statistical reproducibility would be to use another dataset whose characteristics are similar to those of the original dataset (for instance, similar age, sex, scanner distributions). Note that the boundary between statistical and conceptual reproducibility (defined in the next section) is fuzzy. We do not believe it is possible to draw exact frontiers that would delimit the statistical variations that are admissible in a statistical reproducibility study. Finally, it is important that those who conduct the statistical replication study clearly indicate which components of variability they study.

What Does It Require? Here one needs to distinguish between two types of factors: those necessary to *attempt* reproducibility and those that increase the likelihood of *successful* reproducibility.

Regarding the first type, most factors are common with those for exact reproducibility. Code needs to be accessible so that variations coming from reimplementations do not impact the replication. Random seeds, GPU model, or other software/execution parameters will not be set to be identical because the aim is precisely to check if the findings of the study are statistically reproducible under such variations. Knowing their value in the original study is nevertheless useful in order to dissect potential reasons for failed replication. Trained models are in a similar situation: they will usually not be used for statistical replication (models will be retrained) but shall prove useful to dissect potential failures. Data

¹⁵We use the term of [19] although with a slightly different (more extensive) meaning.

accessibility is also very valuable because it will allow studying different data splits, or subsamples.

The abovementioned elements make it possible for other researchers to attempt statistical replication of a given study. On the other hand, there are features of the original study that will make such replication more likely to be successful (equivalently, one could say that the original findings are robust). One important factor is that the original study reports error margins (reporting the standard error or equivalently a confidence interval). It is important in this specific context because statistical reproducibility does not aim at obtaining (and cannot obtain) exactly the same results. One wants the results to be *compatible* with original ones: typically a successful replication would produce results which are within the error margin of the original study. Beyond the topic of statistical reproducibility, the report of error margins is of great importance in general, in particular in the field of ML for medical imaging, because it provides a precision on the estimates of the performance. Unfortunately, this practice is still too uncommon in the ML field as a whole [19]. Even worse, it is not uncommon to find faulty interpretations of estimates. For instance, one should never estimate standard errors (SE) from multiple runs of a cross-validation, as the number of runs can be made arbitrarily large and as a consequence the SE arbitrarily small (*see* [4]). A very common example is papers which report empirical standard deviation (SD) across k -folds (or more generally across splits). Unlike what is quite widely believed, this value does not allow to gauge the precision of the performance estimation. It provides some insight on the variability of the learning procedure under variations of the training and validation sets. Further, keep in mind that when *the number of splits* is small, such gauge will be very rough. *When the number of splits is sufficiently large* (and typically using random splits rather than k -fold), it is possible to assess if a “learner” (i.e., an ML procedure to perform a task) is superior to another one by counting the fraction of folds on which it obtains superior performance (e.g., 75%) [53]. *See* Chap. 21 for more details on this question. However, in no case can such procedures estimate the precision of the performance of the trained model, in other words the precision of the computed biomarker or computer-aided diagnosis tool. This requires an independent test set, from which SE and confidence intervals can be computed.

Why Is It Useful? Statistical replication has many merits. First, by reassessing ML methods using different data splits, one can spot faulty procedures including data leakage which is prevalent in the field of medical imaging [54–57]. *See* refs. 4, 5 for more details on data leakage. Beyond procedures which are clearly wrong, it can also detect lack of robustness to different parameters. One would

consider that the procedure is not statistically replicable if it leads to substantially different results under different train/test data splits, different random seeds, or small changes in hyperparameters. Such an ML algorithm would display poor robustness and would be unlikely to be of future clinical use. Note that, regarding the use of different train/test data splits, these would need to preserve a distribution of metadata (for instance, age, sex, diagnosis...) between train and test that is similar to that of the original study. Most classically, if the original study has stratified the splits, the statistical replication study would also need to stratify the splits. Using different distributions (e.g., not stratified) is also interesting but, in our view, falls within conceptual rather than statistical reproducibility. Furthermore, it is very interesting to attempt replication on a different dataset with statistically equivalent characteristics: for instance, another subsample which has not been used in the original study (but comes from the same larger dataset) or a different dataset but with similar characteristics (e.g., same MRI scanners, similar age, similar disease stage...). Unsuccessful replication may be an indication of overfitting of the dataset of the original study through excessive experimentation with different architectures or hyperparameters which ended up with a method that would work only on this very specific dataset. This is referred to as the *researcher degrees of freedom* [6, 22]. This concept extends beyond the field of ML. It actually comes from experimental sciences where different statistical procedures are tried until a statistically significant result is found, a bad practice known as *p-hacking* [58]. It is important that researchers in our field have this problem in mind. Experimental sciences have proposed *preregistered* and *registered* studies as a potential solution to ban such bad practices. Preregistration means that the research plan is written down and made public before the study starts. It can, for example, be published on the Open Science Framework website.¹⁶ This mechanism reduces the researcher degrees of freedom and is thus likely to lead to more robust results. Registration goes one step further. The research plan is submitted to a journal and peer-reviewed. Thus (most of) the peer review is done before the results are known. It has the additional advantage of putting more focus on methodological soundness than on the groundbreaking nature of results (for instance, negative results will be published). More details about preregistration and registration can be found in [59]. Preregistration and registration are not yet widely used in ML for medical imaging. Such practices would certainly not fit all studies because they leave no room for methodological creativity. On the other hand, they should be very valuable to experimental

¹⁶ <https://osf.io>.

studies aiming at validating ML methods. We believe that, as a community, we should try to adapt such procedures to our field.

5 Conceptual Reproducibility

What Is It? Conceptual reproducibility can be seen as the ultimate goal: the one which lead to the consolidation of scientific knowledge. The general idea is to be able to validate the findings under conceptually similar conditions.¹⁷ Conceptually similar means that the method, the data, and the experiments are compatible with the claims of the original study but they are not identical. We will come back to the notion of claims of a study, and their relationships to generalizability and limitations, later in this section.

What Does It Require? Again, we may distinguish between factors that make it possible to *attempt* replication and those that will make it more likely to be *successful*.

In theory, nothing but the original paper should be strictly necessary. Nevertheless, this assumes that the original paper has adhered to the scientific gold standard of providing all details necessary for replication: not only a description of the methods which makes reimplementation possible but a detailed description of the datasets and experimental procedure. It is particularly worrisome that many medical imaging publications do not even report basic demographic statistics [30]. [14] argues that the replication should be independent of the implementation. We agree in principle but believe that such requirement would considerably lower the number of conceptual replication attempts, while more are needed to advance our field in a steadier manner. In practice, it is extremely useful to be able to access the code, not only to save a lot of time but also to make sure that an unsuccessful replication is not due to a faulty reimplementation. The same can be said for trained models. Access to the original data can be useful to dissect the potential reasons for differences in results. In summary, none of the elements of exact reproducibility are required, all of them are welcome.

There are several characteristics of an original study that make it less likely for it to be replicated. Low sample size not only means that it is less likely to find a true effect if it exists but also increases the odds that a positive finding is false [9]. This is not only true in ML but in experimental sciences in general. Ideally, the sample size should be justified by a previous power analysis [24]. Causes for failure of statistical reproducibility also apply here. In particular, too many researcher degrees of freedom increase the likelihood of

¹⁷ Again, we use the term of [19] although with a slightly different meaning.

having built a method that is overly specific to a dataset. Another problem is that the datasets used in medical imaging ML papers are very often not representative of what would be found in the clinic [30]. Indeed, they often come from research datasets where the inclusion criteria are specific, the medical imaging protocols are harmonized, and the data quality is controlled. Thus, it is necessary to have more studies including clinical routine data (e.g., [60, 61]). Finally, it is very important to have in mind that most scientific findings will not universally replicate but that the replication will only succeed under specific conditions. This is why it is critical that scientific papers precisely define their claims and their limitations. For instance, a claim could be that a given algorithm can segment brain tumors with a Dice of 0.9 ± 0.02 when the MR images are acquired at 3 Tesla and have only minimal artifacts. The same paper would mention as limitations that it is unclear how the algorithm would perform at 1.5 Tesla or with data of lower quality. One can see that stating clear claims and limitations will allow defining the scope of conceptual replication studies. Studies outside that scope would aim at studying generalizability beyond original claims.

Why Is It Useful? As mentioned above, conceptual reproducibility is the ultimate goal, the one which, through accumulation of evidence, builds consensus about new scientific knowledge. Its utility in general is thus obvious. More specifically, it provides different benefits. In particular, in the field of ML for medical imaging, it allows studying the generalizability of a method. It is thus a step towards its applicability to the clinic. To that aim, the use of multiple datasets is of paramount importance. This will not only allow ruling out that a method is overly specific to a given dataset. It will allow defining which are the bounds within which the method applies. This includes the machine model, the acquisition parameters, and the data quality. It also includes factors which are unrelated to imaging such as population age, sex, geographic origin, disease severity, and others.

6 Measurement Reproducibility

What Is It? Measurement reproducibility is the study of the variability of a specific measurement under different acquisition conditions. We are aware that, at first sight, this concept does not fit ideally in our taxonomy (*see* Subheading 7.1 for a more detailed discussion). Nevertheless, we chose to present it as a separate entity because this is a very common meaning of the word reproducibility

in medical imaging¹⁸ (e.g., [62–69]) and we thus believe that it deserves a special treatment. Here, we consider an algorithm that produces a measurement for each individual patient (for instance, the volume of an anatomical structure computed by a segmentation method). A prototypical example of measurement reproducibility is the test-retest reproducibility: how much does the measure vary when applied to two different scans of the same patient? One can then introduce different variations: scans on the same day or not, scans on the same or different machines, systematic addition of noise or artifacts to the data.... Finally, some authors call inter-method reproducibility the comparison of different software packages for the measurement of the same anatomical entity [70]. We do not believe this falls within the topic of reproducibility but rather of methods' comparison.¹⁹

What Does It Require? The code is necessary to make sure that variations do not depend on implementation and to ease the reproducibility study. The trained models are also very welcome to facilitate the process. It is then necessary to have access to test-retest data, meaning different acquisitions of the same patient. As mentioned above: the more varied these different acquisitions, the more extensive the study. Ideally, one would want to have access to scans performed on the same day [62, 67], on different days [65, 66], at different times during the day (e.g., before or after caffeine consumption, a factor which affects functional MRI measures [71]), on different imaging devices [63], and with different acquisition parameters [68]...It is unlikely to obtain that many scans for the same patients. A more feasible approach is to study these different factors of variations for different patients. Furthermore, starting with a given image, it is possible to simulate different types of alterations and defects by adding them to the original image. This can be very useful because it allows generating very large numbers of images easily and to control for specific imaging characteristics (such as, e.g., the level of noise or the strength of motion artifacts). Such simulations can involve completely synthetic images called phantoms [72] which mimic real images. It can also be done through the addition of defects to real images [73–75]. Ideally, measurement reproducibility should be performed in different populations of participants separately (for instance, a child with autism spectrum disorder or a patient with Parkinson's disease is more likely to move during the acquisition, and the image is thus more likely to be affected by motion artifacts).

¹⁸ Note that the word is used to evaluate reproducibility of automatic methods across different scans of the same subject but also when a human rater is involved (manual or semi-automated measurements), including intra-rater (measurement twice by the same rater) and inter-rater (two different raters) reproducibility from a single scan.

¹⁹ It is of interest to compare which of them is the most accurate or robust, with respect to a ground truth. However, as mentioned above, we do not believe it falls within the topic of reproducibility.

Why Is It Useful? Measurement reproducibility is central for measurement sciences, and medical imaging is one of those. It is an extremely precious information to the user (for instance, the radiologist). Indeed, it provides, at the individual patient level, and ideally for different categories of patients, the precision that they may expect from the measurement tool. There is a wide tradition to perform such reproducibility studies in radiology journals. We believe that it would be very welcome that it becomes more commonplace in the ML for medical imaging community.

7 Discussion

7.1 About the Different Types of Reproducibility

We have presented different types of reproducibility. Our taxonomy is not original nor aims at being universal. The boundaries between types are partly fuzzy. For instance, to which degree replication with a different but similar dataset should be considered *statistical* or *conceptual* reproducibility? We do not believe such questions to be of great importance. Rather, it is fruitful, following Gundersen and Kjensmo [14] and Peng [76], to consider reproducibility as a spectrum. In particular, one can consider that the first three types provide increasing support for a finding: conceptual provides more support than statistical which in turns provides more support than exact. The amount of components necessary to perform them is in the reverse order: exact requires more than statistical which requires more than conceptual. Does it mean that only conceptual reproducibility matters? Absolutely not. As we mentioned, other types of reproducibility are necessary to dissect why a given replication has failed as well as to better specify the bounds within which a scientific claim is valid. Last but not least, exact reproducibility also helps build trust in science.

We must admit that measurement reproducibility does not fit very well in this landscape. Moreover, one could also argue that it is a type of conceptual reproducibility, which is partly true as it aims at studying the reproducibility when varying the input data. We nevertheless believe it deserves a special treatment, for several reasons. First, here reproducibility is studied at the individual (i.e., patient) level and not at the population level. Also, the emphasis is on the measurement rather than the finding. Even if it has its role in the building of scientific knowledge, it has specific practical implications for the user. Moreover, as mentioned above, this is actually the most widely used meaning of reproducibility in medical imaging, and it seemed important that the reader is acquainted with it.

7.2 The Many Benefits of Reproducibility

“Der Weg ist das Ziel” is a German saying which can be roughly translated as: “the path is the goal.” Indeed, reproducibility allows researchers to discover many new places down the road before reaching the final destination. Even if this destination is never reached, the benefits of the travel are of major importance. Let us try to list some of them.

There are many individual benefits for researchers and labs. An important one is that aiming at reproducible research results in *reusable* research artifacts. How agreeable it is for a researcher to easily reuse an old code for a new project! How useful it is for a research lab to have data organized according to community standards making it easier to reuse and share! Moreover, papers that come with shared data [22, 77, 78] or code attract [79], on average, more citations. Thus aiming at reproducibility is also in the researchers’ self-interest.

There are also considerable benefits for the scientific community as a whole. As mentioned before, reproducible research is often associated to open code, open data, and available trained models. This allows researchers not only to use them to perform replication studies but also to use these research artifacts for completely different purposes such as building new methods or conducting analysis on pooled datasets. In the specific case of ML for medical imaging, it also allows assessing independently the influence of preprocessing, feature extraction, and ML method. This is particularly important when claims of superiority of a new ML method are made, but the original paper uses overly specific preprocessing steps.

Of course, at the end of the path, the goal itself brings many benefits. These have already largely described in the previous sections so we will just mention them briefly. Conceptual replication studies are necessary for corroborating findings and thus building new scientific knowledge. Statistical replication allows ensuring that results are not due to cherry picking. Exact replication allows detecting errors and increases trust in science in general.

7.3 Awareness Is Rising

Throughout this chapter, we have referred to numerous papers, resources, and tools that demonstrate that awareness regarding reproducibility has strongly risen in the past years.

Various papers and studies have highlighted the lack of reproducibility in different fields (e.g., [11, 15, 19]). In machine learning for medical imaging, Simko et al. [21] have studied the reproducibility of methods (mainly code availability and usability thus restricted to exact reproducibility) published at the Medical Imaging with Deep Learning (MIDL) conference from 2018 to 2022

and found that about 20% of papers came with a repository that was deemed reproducible.

Various papers have been published providing advice and guidelines [10, 17, 22–27]. Some of the guidelines include reproducibility checklists. Some checklists are associated to a specific journal or conference and are provided to the reviewers so that they can take these aspects in consideration when evaluating papers. One can cite, for example, the MICCAI (Medical Image Computing and Computer-Assisted Intervention conference) reproducibility checklist.^{20,21}

Finally, it is very important that reproducibility studies, assessing all aspects of reproducibility (exact, statistical, conceptual, measurement), are performed, published, and widely read. Unfortunately, it is still easier to publish in a high-impact journal a study that is not reproducible but describes exciting results than a replication study. The good news is that this is starting to change. Reproducibility challenges have been proposed in various fields including machine learning²² and medical image computing [80]. In the field of neuroimaging, the journal NeuroImage: Reports publishes Open Data Replication Reports.²³ the Organization for Human Brain Mapping has a replication award.²⁴ and the MRITogether workshop²⁵ emphasizes reproducibility.

7.4 One Size Does Not Fit All

We hope the reader is now convinced of the benefits of aiming towards reproducible research. Does it mean that reproducibility requirements should be the same for all studies? We strongly believe the opposite. To take an extreme example, requiring all studies to be exactly reproducible with minimal efforts (like with running a single command) would be an awful idea. We believe, on the contrary, that reproducibility efforts should vary according to many factors including the type of study and the context in which it is performed. One would probably not have the same level of requirement for a methodological paper and for an extensive medical application with strong claims about clinical applicability. For the former, one may be satisfied with an experiment on a single or a few datasets. For the later, one would expect the study to include

²⁰ <https://miccai2021.org/files/downloads/MICCAI2021-Reproducibility-Checklist.pdf>.

²¹ <https://github.com/JunMa11/MICCAI-Reproducibility-Checklist>.

²² <https://paperswithcode.com/rc2022>.

²³ https://www.journals.elsevier.com/neuroimage-reports/infographics/neuroimage-reports-presents-open-data-replication-reports?utm_campaign=STMJ_176479_SC&utm_medium=email&utm_acid=268008024&SIS_ID=&dgcid=STMJ_176479_SC&CMX_ID=&utm_in=DM292849&utm_source=AC_.

²⁴ <https://www.humanbrainmapping.org/i4a/pages/index.cfm?pageid=3731>.

²⁵ <https://mritogether.esmrb.org/>.

multiple datasets with varying characteristics and a comprehensive assessment of generalizability under different factors such as imaging devices and acquisition parameters. Also, there are some cases where sharing the code is not desired (e.g., because an industrial application is foreseen) or where the code will not adhere to best development practices because it is just a prototype to test a new methodology. Nevertheless, sharing weakly documented code is always better than no sharing at all. Similarly, there are cases where data sharing is difficult or even impossible due to regulatory constraints. As mentioned above, reproducibility is a spectrum. Where a given study should lie in this spectrum should depend on the type of study and the constraints the researchers face.

We thus advocate for a nondogmatic approach to reproducibility. Guidelines are extremely useful, but they should not be carved in stone. Also, we believe that the requirements should be assessed by the reviewers on a case-by-case basis. Indeed, what matters is that the reproducibility level matches the claims made in the paper. Of course, it is a good thing that journals and conferences provide requirements for reporting essential information. It is helpful to researchers and makes the community progress towards better science. Also, some bad practices such as data leakage or p-hacking need to be banished. But we believe that very high reproducibility requirements (e.g., requiring that exact reproducibility is feasible) at the level of a given journal or conference would be counterproductive. Finally, we like the idea of a badging system [27] which would tag papers according to their reproducibility level. It remains to be seen how such system should be implemented.

To conclude, we firmly believe that it is essential for researchers and students in the field of ML for medical imaging to be trained to the concepts and practice of reproducibility. It will be beneficial to them as well as to the community in general. But this does not mean that researchers should aim at perfect reproducibility in all their studies. Diversity in research approaches and practices is also a factor that drives science forward and which should be preserved.

Acknowledgements

The authors are grateful to G. Varoquaux for pointing them towards useful references. This work was supported by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA

Institut Hospitalo-Universitaire-6). ETS acknowledges funding from the 4TU Precision Medicine program supported by High Tech for a Sustainable Future, a framework commissioned by the four Universities of Technology of the Netherlands.

References

1. Seab J, Jagust W, Wong S, Roos M, Reed BR, Budinger T (1988) Quantitative NMR measurements of hippocampal atrophy in Alzheimer's disease. *Magn Reson Med* 8(2):200–208
2. Lehericy S, Baulac M, Chiras J, Pierot L, Martin N, Pillon B, Deweer B, Dubois B, Marsault C (1994) Amygdalohippocampal MR volume measurements in the early stages of Alzheimer disease. *Am J Neuroradiol* 15(5): 929–937
3. Jack CR, Petersen RC, Xu YC, Waring SC, O'Brien PC, Tangalos EG, Smith GE, Ivnik RJ, Kokmen E (1997) Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology* 49(3):786–794
4. Varoquaux G, Colliot O (2022) Evaluating machine learning models and their diagnostic value. HAL preprint hal-03682454. <https://hal.archives-ouvertes.fr/hal-03682454/>
5. Thibeau-Sutre E, Diaz M, Hassanaly R, Routier A, Dormont D, Colliot O, Burgos N (2022) ClinicaDL: an open-source deep learning software for reproducible neuroimaging processing. *Comput Methods Prog Biomed* 220:106818
6. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22: 1359–1366
7. Baker M (2016) 1,500 scientists lift the lid on reproducibility. *Nature* 533:452–454
8. Gundersen OE (2020) The reproducibility crisis is real. *AI Mag* 41(3):103–106
9. Ioannidis JP (2005) Why most published research findings are false. *PLoS Med* 2(8): e124
10. Begley CG, Ioannidis JP (2015) Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res* 116(1): 116–126
11. Collaboration OS (2015) Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716
12. Begley CG (2013) An unappreciated challenge to oncology drug discovery: pitfalls in preclinical research. *Am Soc Clin Oncol Educ Book* 33(1):466–468
13. Sonnenburg S, Braun ML, Ong CS, Bengio S, Bottou L, Holmes G, LeCunn Y, Muller KR, Pereira F, Rasmussen CE et al (2007) The need for open source software in machine learning. *J Mach Learn Res* 8(81):2443–2466
14. Gundersen OE, Kjensmo S (2018) State of the art: reproducibility in artificial intelligence. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32
15. Hutson M (2018) Artificial intelligence faces reproducibility crisis. *Science* 359(6377): 725–726
16. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Waldron L, Wang B, McIntosh C, Goldenberg A, Kundaje A, Greene CS, et al (2020) Transparency and reproducibility in artificial intelligence. *Nature* 586(7829):E14–E16
17. Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d'Alché Buc F, Fox E, Larochelle H (2021) Improving reproducibility in machine learning research: a report from the neurips 2019 reproducibility program. *J Mach Learn Res* 22:1–20
18. Stupple A, Singerman D, Celi LA (2019) The reproducibility crisis in the age of digital medicine. *NPJ Digit Med* 2(1):1–3
19. McDermott M, Wang S, Marinsek N, Ranganath R, Ghassemi M, Foschini L (2019) Reproducibility in machine learning for health. arXiv preprint arXiv:190701463
20. Beam AL, Manrai AK, Ghassemi M (2020) Challenges to the reproducibility of machine learning models in health care. *JAMA* 323(4): 305–306
21. Simko A, Garpebring A, Jonsson J, Nyholm T, Löfstedt T (2022) Reproducibility of the methods in medical imaging with deep learning. arXiv preprint arXiv:221011146
22. Gorgolewski KJ, Poldrack RA (2016) A practical guide for improving transparency and reproducibility in neuroimaging research. *PLoS Biol* 14(7):e1002506
23. Nichols TE, Das S, Eickhoff SB, Evans AC, Glatard T, Hanke M, Kriegeskorte N, Milham MP, Poldrack RA, Poline JB et al (2017) Best practices in data analysis and sharing in neuroimaging using MRI. *Nat Neurosci* 20(3): 299–303

24. Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, Nichols TE, Poline JB, Vul E, Yarkoni T (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci* 18(2):115–126
25. Niso G, Botvinik-Nezer R, Appelhoff S, De La Vega A, Esteban O, Etzel JA, Finc K, Ganz M, Gau R, Halchenko YO et al (2022) Open and reproducible neuroimaging: from study inception to publication. *Neuroimage* 263:119623
26. Turkyilmaz-van der Velden Y, Dintzner N, Teperek M (2020) Reproducibility starts from you today. *Patterns* 1(6):100099
27. Heil BJ, Hoffman MM, Markowitz F, Lee SI, Greene CS, Hicks SC (2021) Reproducibility standards for machine learning in the life sciences. *Nat Methods* 18(10):1132–1135
28. Varoquaux G (2018) Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* 180:68–77
29. Button KS, Ioannidis J, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14(5):365–376
30. Varoquaux G, Cheplygina V (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med* 5(1):1–8
31. Bouthillier X, Laurent C, Vincent P (2019) Unreproducible research is reproducible. In: International Conference on Machine Learning, PMLR, pp 725–734
32. Langer SG, Shih G, Nagy P, Landman BA (2018) Collaborative and reproducible research: goals, challenges, and strategies. *J Digit Imaging* 31(3):275–282
33. Goodman SN, Fanelli D, Ioannidis JP (2016) What does research reproducibility mean? *Sci Transl Med* 8(341):341ps12–341ps12
34. Plesser HE (2018) Reproducibility vs. replicability: a brief history of a confused terminology. *Front Neuroinfor* 11:76
35. McDermott MB, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M (2021) Reproducibility in machine learning for health research: still a ways to go. *Sci Transl Med* 13(586):eabb1655
36. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, et al (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 3(1):1–9
37. Gabelica M, Bojčić R, Puljak L (2022) Many researchers were not compliant with their published data sharing statement: mixed-methods study. *J Clin Epidemiol* 150:33–41
38. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, et al (2016) The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* 3(1):1–9
39. Bourget MH, Kamentsky L, Ghosh SS, Mazzamuto G, Lazari A, Markiewicz CJ, Oostenveld R, Niso G, Halchenko YO, Lipp I, et al (2022) Microscopy-BIDS: an extension to the Brain imaging data structure for microscopy data. *Front Neurosci* 16: e871228
40. Saborit-Torres J, Saenz-Gamboa J, Montell J, Salinas J, Gómez J, Stefan I, Caparrós M, García-García F, Domenech J, Manjón J, et al (2020) Medical imaging data structure extended to multiple modalities and anatomical regions. arXiv preprint arXiv:201000434
41. Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert MO, Chupin M, Benali H, Colliot O (2011) Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *Neuroimage* 56(2): 766–781
42. Samper-González J, Burgos N, Bottani S, Fontanella S, Lu P, Marcoux A, Routier A, Guillou J, Bacci M, Wen J, et al (2018) Reproducible evaluation of classification methods in Alzheimer's disease: framework and application to MRI and PET data. *Neuroimage* 183:504–521
43. Karakuzu A, DuPre E, Tetrel L, Bermudez P, Boudreau M, Chin M, Poline JB, Das S, Bellec P, Stikov N (2022) Neurolibre: a preprint server for full-fledged reproducible neuroscience. OSF Preprints
44. Crane M (2018) Questionable answers in question answering research: reproducibility and variability of published results. *Trans Assoc Comput Linguist* 6:241–252
45. Jungo A, Scheidegger O, Reyes M, Balsiger F (2021) pymia: a python package for data handling and evaluation in deep learning-based medical image analysis. *Comput Methods Program Biomed* 198:105796. <https://doi.org/10.1016/j.cmpb.2020.105796>
46. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson U, et al (2021) Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21), pp 2633–2650

47. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016) Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp 308–318
48. Gorgolewski KJ, Alfaro-Almagro F, Auer T, Bellec P, Capotă M, Chakravarty MM, Churchill NW, Cohen AL, Craddock RC, Devenyi GA, et al (2017) BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLoS Comput Biol* 13(3):e1005209
49. Routier A, Burgos N, Díaz M, Bacci M, Bottani S, El-Rifai O, Fontanella S, Gori P, Guillon J, Guyot A, et al (2021) Clinica: an open-source software platform for reproducible clinical neuroscience studies. *Front Neuroinform* 15:e689675
50. McCormick M, Liu X, Jomier J, Marion C, Ibanez L (2014) ITK: enabling reproducible research and open science. *Front Neuroinform* 8:13
51. Yoo TS, Ackerman MJ, Lorensen WE, Schroeder W, Chalana V, Aylward S, Metaxas D, Whitaker R (2002) Engineering and algorithm design for an image processing API: a technical report on ITK—the insight toolkit. In: Medicine Meets Virtual Reality 02/10, IOS press, pp 586–592
52. Drummond C (2009) Replicability is not reproducibility: nor is it good science. In: Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML, vol 1
53. Bouthillier X, Delaunay P, Bronzi M, Trofimov A, Nichyporuk B, Szeto J, Mohammadi Sepahvand N, Raff E, Madan K, Voleti V, et al (2021) Accounting for variance in machine learning benchmarks. *Proc Mach Learn Syst* 3:747–769
54. Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Dormont D, Durrleman S, Burgos N, Colliot O (2020) Convolutional neural networks for classification of Alzheimer’s disease: overview and reproducible evaluation. *Med Image Anal* 63:101694
55. Samala RK, Chan HP, Hadjiiski L, Koneru S (2020) Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks. In: Proceedings of SPIE Medical Imaging 2020: Computer-Aided Diagnosis, International Society for Optics and Photonics, vol 11314, p 1131416
56. Panwar H, Gupta PK, Siddiqui MK, Morales-Menendez R, Singh V (2020) Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. *Chaos, Solitons Fractals* 138:109944
57. Bussola N, Marcolini A, Maggio V, Jurman G, Furlanello C (2021) AI slipping on tiles: Data leakage in Digital Pathology. In: Del Bimbo A, Cucchiara R, Sclaroff S, Farinella GM, Mei T, Bertini M, Escalante HJ, Vezzani R (eds) Pattern recognition. ICPR International Workshops and Challenges, Springer International Publishing, Cham. Lecture notes in computer science, pp 167–182. https://doi.org/10.1007/978-3-030-68763-2_13
58. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The extent and consequences of p-hacking in science. *PLoS Biol* 13(3):e1002106
59. Henderson EL (2022) A guide to preregistration and registered reports. Preprint. <https://osf.io/preprints/metaarxiv/x7aq/> download
60. Bottani S, Burgos N, Maire A, Wild A, Ströer S, Dormont D, Colliot O, Group AS et al (2022) Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse. *Med Image Anal* 75: 102219
61. Perkuhn M, Stavrinou P, Thiele F, Shakirin G, Mohan M, Garmpis D, Kabbasch C, Borggrefe J (2018) Clinical evaluation of a multiparametric deep learning model for glioblastoma segmentation using heterogeneous magnetic resonance imaging data from clinical routine. *Investig Radiol* 53(11):647
62. Lukas C, Hahn HK, Bellenberg B, Rexilius J, Schmid G, Schimrigk SK, Przuntek H, Köster O, Peitgen HO (2004) Sensitivity and reproducibility of a new fast 3D segmentation technique for clinical MR-based brain volumetry in multiple sclerosis. *Neuroradiology* 46(11):906–915
63. Borga M, Ahlgren A, Romu T, Widholm P, Dahlqvist Leinhard O, West J (2020) Reproducibility and repeatability of MRI-based body composition analysis. *Magn Reson Med* 84(6): 3146–3156
64. Chard DT, Parker GJ, Griffin CM, Thompson AJ, Miller DH (2002) The reproducibility and sensitivity of brain tissue volume measurements derived from an SPM-based segmentation methodology. *J Magn Reson Imaging: Off J Int Soc Magn Reson Med* 15(3):259–267
65. de Boer R, Vrooman HA, Ikram MA, Vernooij MW, Breteler MM, van der Lugt A, Niessen WJ (2010) Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods. *Neuroimage* 51(3):1047–1056

66. Lemieux L, Hagemann G, Krakow K, Woermann FG (1999) Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data. *Magn Reson Med: Off J Int Soc Magn Reson Med* 42(1):127–135
67. Tudorascu DL, Karim HT, Maronge JM, Alhilali L, Fakhraian S, Aizenstein HJ, Muschelli J, Crainiceanu CM (2016) Reproducibility and bias in healthy brain segmentation: comparison of two popular neuroimaging platforms. *Front Neurosci* 10:503
68. Yamashita R, Perrin T, Chakraborty J, Chou JF, Horvat N, Koszalka MA, Midya A, Gonen M, Allen P, Jarnagin WR et al (2020) Radiomic feature reproducibility in contrast-enhanced CT of the pancreas is affected by variabilities in scan parameters and manual segmentation. *Euro Radiol* 30(1):195–205
69. Poldrack RA, Whitaker K, Kennedy DN (2019) Introduction to the special issue on reproducibility in neuroimaging. *Neuroimage* 218:116357
70. Palumbo L, Bosco P, Fantacci M, Ferrari E, Oliva P, Spera G, Retico A (2019) Evaluation of the intra-and inter-method agreement of brain MRI segmentation software packages: a comparison between SPM12 and FreeSurfer v6.0. *Phys Med* 64:261–272
71. Laurienti PJ, Field AS, Burdette JH, Maldjian JA, Yen YF, Moody DM (2002) Dietary caffeine consumption modulates fMRI measures. *Neuroimage* 17(2):751–757
72. Collins DL, Zijdenbos AP, Kollokian V, Sled JG, Kabani NJ, Holmes CJ, Evans AC (1998) Design and construction of a realistic digital brain phantom. *IEEE Trans Med Imaging* 17(3):463–468
73. Shaw R, Sudre C, Ourselin S, Cardoso MJ (2018) MRI K-space motion artefact augmentation: Model robustness and task-specific uncertainty. In: *Medical Imaging with Deep Learning – MIDL 2018*
74. Duffy BA, Zhang W, Tang H, Zhao L (2018) Retrospective correction of motion artifact affected structural MRI images using deep learning of simulated motion. In: *Medical Imaging with Deep Learning – MIDL 2018*
75. Loizillon S, Bottani S, Maire A, Ströer S, Dormont D, Colliot O, Burgos N (2023) Transfer learning from synthetic to routine clinical data for motion artefact detection in brain t1-weighted MRI. In: *SPIE Medical Imaging 2023: Image Processing*
76. Peng RD (2011) Reproducible research in computational science. *Science* 334(6060): 1226–1227
77. Piwowar HA, Day RS, Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. *PLoS One* 2(3): e308
78. Piwowar HA, Vision TJ (2013) Data reuse and the open data citation advantage. *PeerJ* 1:e175
79. Vandewalle P (2012) Code sharing is associated with research impact in image processing. *Comput Sci Eng* 14(4):42–47
80. Balsiger F, Jungo A, Chen J, Ezhov I, Liu S, Ma J, Paetzold JC, Sekuboyina A, Shit S, Suter Y et al (2021) The miccai hackathon on reproducibility, diversity, and selection of papers at the miccai conference. *arXiv preprint arXiv:210305437*

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third-party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 22

Interpretability of Machine Learning Methods Applied to Neuroimaging

Elina Thibeau-Sutre, Sasha Collin, Ninon Burgos, and Olivier Colliot

Abstract

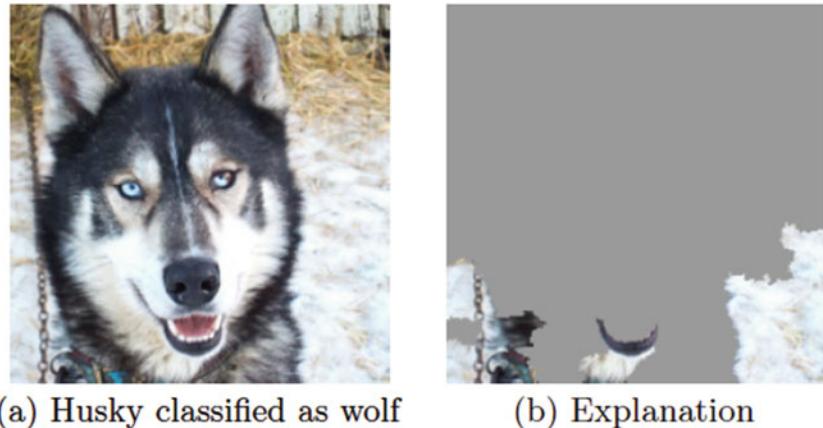
Deep learning methods have become very popular for the processing of natural images and were then successfully adapted to the neuroimaging field. As these methods are non-transparent, interpretability methods are needed to validate them and ensure their reliability. Indeed, it has been shown that deep learning models may obtain high performance even when using irrelevant features, by exploiting biases in the training set. Such undesirable situations can potentially be detected by using interpretability methods. Recently, many methods have been proposed to interpret neural networks. However, this domain is not mature yet. Machine learning users face two major issues when aiming to interpret their models: which method to choose and how to assess its reliability. Here, we aim at providing answers to these questions by presenting the most common interpretability methods and metrics developed to assess their reliability, as well as their applications and benchmarks in the neuroimaging context. Note that this is not an exhaustive survey: we aimed to focus on the studies which we found to be the most representative and relevant.

Key words Interpretability, Saliency, Machine learning, Deep learning, Neuroimaging, Brain disorders

1 Introduction

1.1 Need for Interpretability

Many metrics have been developed to evaluate the performance of machine learning (ML) systems. In the case of supervised systems, these metrics compare the output of the algorithm to a ground truth, in order to evaluate its ability to reproduce a label given by a physician. However, the users (patients and clinicians) may want more information before relying on such systems. On which features is the model relying to compute the results? Are these features close to the way a clinician thinks? If not, why? This questioning coming from the actors of the medical field is justified, as errors in real life may lead to dramatic consequences. Trust into ML systems cannot be built only based on a set of metrics evaluating the performance of the system. Indeed, various examples of machine learning systems taking correct decisions for the wrong reasons



(a) Husky classified as wolf (b) Explanation

Fig. 1 Example of an interpretability method highlighting why a network took the wrong decision. The explained classifier was trained on the binary task “Husky” vs “Wolf.” The pixels used by the model are actually in the background and highlight the snow. (Adapted from [1]. Permission to reuse was kindly granted by the authors)

exist, e.g., [1–3]. Thus, even though their performance is high, they may be unreliable and, for instance, not generalize well to slightly different data sets. One can try to prevent this issue by interpreting the model with an appropriate method whose output will highlight the reasons why a model took its decision.

In [1], the authors show a now classical case of a system that correctly classifies images for wrong reasons. They purposely designed a biased data set in which wolves always are in a snowy environment whereas huskies are not. Then, they trained a classifier to differentiate wolves from huskies: this classifier had good accuracy but classified wolves as huskies when the background was snowy and huskies as wolves when there was no snow. Using an interpretability method, they further highlighted that the classifier was looking at the background and not at the animal (*see* Fig. 1).

Another study [2] detected a bias in ImageNet (a widely used data set of natural images) as the interpretation of images with the label “chocolate sauce” highlighted the importance of the spoon. Indeed, ImageNet “chocolate sauce” images often contained spoons, leading to a spurious correlation. There are also examples of similar problems in medical applications. For instance, a recent paper [3] showed with interpretability methods that some deep learning systems detecting COVID-19 from chest radiographs actually relied on confounding factors rather than on the actual pathological features. Indeed, their model focused on other regions than the lungs to evaluate the COVID-19 status (edges, diaphragm, and cardiac silhouette). Of note, their model was trained on public data sets which were used by many studies.

1.2 How to Interpret Models

According to [4], model interpretability can be broken down into two categories: transparency and post hoc explanations.

A model can be considered as transparent when it (or all parts of it) can be fully understood as such, or when the learning process is understandable. A natural and common candidate that fits, at first sight, these criteria is the linear regression algorithm, where coefficients are usually seen as the individual contributions of the input features. Another candidate is the decision tree approach where model predictions can be broken down into a series of understandable operations. One can reasonably consider these models as transparent: one can easily identify the features that were used to take the decision. However, one may need to be cautious not to push too far the medical interpretation. Indeed, the fact that a feature has not been used by the model does not mean that it is not associated with the target. It just means that the model did not need it to increase its performance. For instance, a classifier aiming at diagnosing Alzheimer's disease may need only a set of regions (for instance, from the medial temporal lobe of the brain) to achieve an optimal performance. This does not mean that other brain regions are not affected by the disease, just that they were not used by the model to take its decision. This is the case, for example, for sparse models like LASSO, but also standard multiple linear regressions. Moreover, features given as input to transparent models are often highly engineered, and choices made before the training step (preprocessing, feature selection) may also hurt the transparency of the whole framework. Nevertheless, in spite of these caveats, such models can reasonably be considered transparent, in particular when compared to deep neural networks which are intrinsically black boxes.

The second category of interpretability methods, post hoc interpretations, allows dealing with non-transparent models. Xie et al. [5] proposed a taxonomy in three categories: *visualization* methods consist in extracting an attribution map of the same size as the input whose intensities allow knowing where the algorithm focused its attention, *distillation* approaches consist in reproducing the behavior of a black box model with a transparent one, and *intrinsic* strategies include interpretability components within the framework, which are trained along with the main task (e.g., a classification). In the present work, we focus on this second category of methods (post hoc) and proposed a new taxonomy including other methods of interpretation (see Fig. 2). Post hoc interpretability is the most used category nowadays, as it allows interpreting deep learning methods that became the state of the art for many tasks in neuroimaging, as in other application fields.

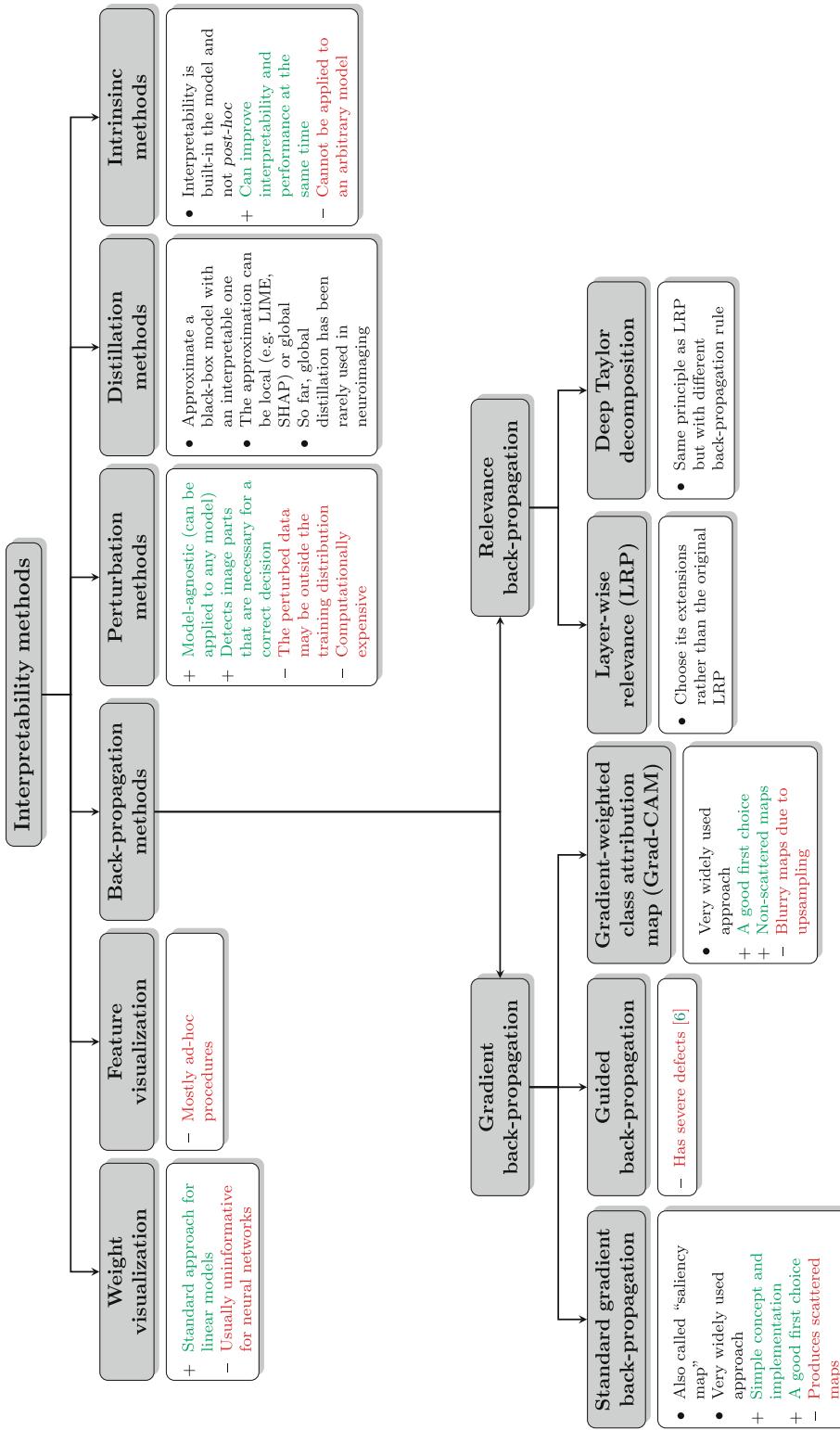


Fig. 2 Taxonomy of the main interpretability methods

1.3 Chapter Content and Outline

This chapter focuses on methods developed to interpret non-transparent machine learning systems, mainly deep learning systems, computing classification, or regression tasks from high-dimensional inputs. The interpretability of other frameworks (in particular generative models such as variational autoencoders or generative adversarial networks) is not covered as there are not enough studies addressing them. It may be because high-dimensional outputs (such as images) are easier to interpret “as such,” whereas small dimensional outputs (such as scalars) are less transparent.

Most interpretability methods presented in this chapter produce an attribution map: an array with the same dimensions as that of the input (up to a resizing) that can be overlaid on top of the input in order to exhibit an explanation of the model prediction. In the literature, many different terms may coexist to name this output such as saliency map, interpretation map, or heatmap. To avoid misunderstandings, in the following, we will only use the term “attribution map.”

The chapter is organized as follows. Subheading 2 presents the most commonly used interpretability methods proposed for computer vision, independently of medical applications. It also describes metrics developed to evaluate the reliability of interpretability methods. Then, Subheading 3 details their application to neuroimaging. Finally, Subheading 4 discusses current limitations of interpretability methods, presents benchmarks conducted in the neuroimaging field, and gives some advice to the readers who would like to interpret their own models.

Mathematical notations and abbreviations used during this chapter are summarized in Tables 1 and 2. A short reminder on neural network training procedure and a brief description of the diseases mentioned in the present chapter are provided in Appendices A and B.

2 Interpretability Methods

This section presents the main interpretability methods proposed in the domain of computer vision. We restrict ourselves to the methods that have been applied to the neuroimaging domain (the applications themselves being presented in Subheading 3). The outline of this section is largely inspired from the one proposed by Xie et al. [5]:

1. **Weight visualization** consists in directly visualizing weights learned by the model, which is natural for linear models but quite less informative for deep learning networks.

2. **Feature map visualization** consists in displaying intermediate results produced by a deep learning network to better understand its operation principle.
3. **Back-propagation methods** back-propagate a signal through the machine learning system from the output node of interest o_c to the level of the input to produce an attribution map.

Table 1
Mathematical notations

X_0 is the input tensor given to the network, and X refers to any input, sampled from the set \mathcal{X} .
y is a vector of target classes corresponding to the input.
f is a network of L layers. The first layer is the closest to the input; the last layer is the closest to the output. A layer is a function.
g is a transparent function which aims at reproducing the behavior of f .
w and b are the weights and the bias associated to a linear function (e.g., in a fully connected layer).
u and v are locations (set of coordinates) corresponding to a node in a feature map. They belong respectively to the set \mathcal{U} and \mathcal{V} .
$A_k^{(l)}(u)$ is the value of the feature map computed by layer l , of K channels at channel k , at position u .
$R_k^{(l)}(u)$ is the value of a property back-propagated through the $l+1$, of K channels at channel k , at position u . $R^{(l)}$ and $A^{(l)}$ have the same number of channels.
o_c is the output node of interest (in a classification framework, it corresponds to the node of the class c).
S_c is an attribution map corresponding to the output node o_c .
m is a mask of perturbations. It can be applied to X to compute its perturbed version X^m .
Φ is a function producing a perturbed version of an input X .
Γ_c is the function computing the attribution map S_c from the black-box function f and an input X_0 .

Table 2
Abbreviations

CAM Class activation maps
CNN Convolutional neural network
CT Computed tomography
Grad-CAM Gradient-weighted class activation mapping
LIME Local interpretable model-agnostic explanations
LRP Layer-wise relevance
MRI Magnetic resonance imaging
SHAP SHapley Additive exPlanations
T1w T1-weighted [Magnetic Resonance Imaging]

4. **Perturbation methods** locally perturb the input and evaluate the difference in performance between using the original input and the perturbed version to infer which parts of the input are relevant for the machine learning system.
5. **Distillation** approximates the behavior of a black box model with a more transparent one and then draw conclusions from this new model.
6. **Intrinsic** includes the only methods of this chapter that are not post hoc explanations: in this case, interpretability is obtained thanks to components of the framework that are trained at the same time as the model.

Finally, for the methods producing an attribution map, a section is dedicated to the metrics used to evaluate different properties (e.g., reliability or human intelligibility) of the maps.

We caution readers that this taxonomy is not perfect: some methods may belong to several categories (e.g., LIME and SHAP could belong either to perturbation or distillation methods). Moreover, interpretability is still an active research field, and then some categories may (dis)appear or be fused in the future.

The interpretability methods were (most of the time) originally proposed in the context of a classification task. In this case, the network outputs an array of size C , corresponding to the number of different labels existing in the data set, and the goal is to know how the output node corresponding to a particular class c interacts with the input or with other parts of the network. However, these techniques can be extended to other tasks: for example, for a regression task, we will just have to consider the output node containing the continuous variable learned by the network. Moreover, some methods do not depend on the nature of the algorithm (e.g., standard perturbation or LIME) and can be applied to any machine learning algorithm.

2.1 Weight Visualization

At first sight, one of can be tempted to directly visualize the weights learned by the algorithm. This method is really simple, as it does not require further processing. However, even though it can make sense for linear models, it is not very informative for most networks unless they are specially designed for this interpretation.

This is the case for AlexNet [7], a convolutional neural network (CNN) trained on natural images (ImageNet). In this network the size of the kernels in the first layer is large enough (11×11) to distinguish patterns of interest. Moreover, as the three channels in the first layer correspond to the three color channels of the images (red, green, and blue), the values of the kernels can also be represented in terms of colors (this is not the case for hidden layers, in which the meaning of the channels is lost). The 96 kernels of the first layer were illustrated in the original article as in Fig. 3.

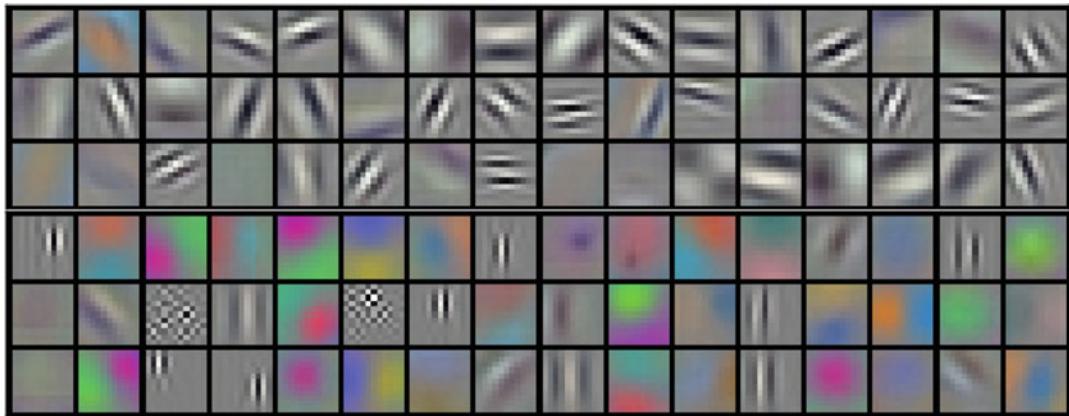


Fig. 3 96 convolutional kernels of size $3 \times 11 \times 11$ learned by the first convolutional layer on the $3 \times 224 \times 224$ input images by AlexNet. (Adapted from [7]. Permission to reuse was kindly granted by the authors)

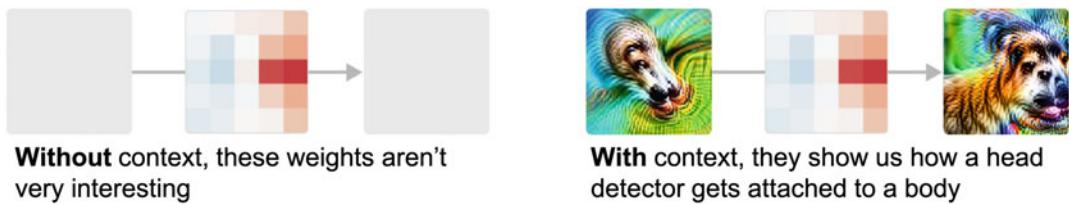


Fig. 4 The weights of small kernels in hidden layers (here 5×5) can be really difficult to interpret alone. Here some context allows better understanding how it modulates the interaction between concepts conveyed by the input and the output. (Adapted from [8] (CC BY 4.0))

However, for hidden layers, this kind of interpretation may be misleading as nonlinearity activation layers are added between the convolutions and fully connected layers; this is why they only visualized the weights of the first layer.

To understand the weight visualization in hidden layers of a network, Voss et al. [8] proposed to add some context to the input and the output channels. This way they enriched the weight visualization with feature visualization methods able to generate an image corresponding to the input node and the output node (*see Fig. 4*). However, the feature visualization methods used to bring some context can also be difficult to interpret themselves, and then it only moves the interpretability problem from weights to features.

2.2 Feature Map Visualization

Feature maps are the results of intermediate computations done from the input and resulting in the output value. Then, it seems natural to visualize them or link them to concepts to understand how the input is successively transformed into the output.

Methods described in this section aim at highlighting which concepts a feature map (or part of it) A conveys.

2.2.1 Direct Interpretation

The output of a convolution has the same shape as its input: a 2D image processed by a convolution will become another 2D image (the size may vary). Then, it is possible to directly visualize these feature maps and compare them to the input to understand the operations performed by the network. However, the number of filters of convolutional layers (often a hundred) makes the interpretation difficult as a high number of images must be interpreted for a single input.

Instead of directly visualizing the feature map A , it is possible to study the latent space including all the values of the samples of a data set at the level of the feature map A . Then, it is possible to study the deformations of the input by drawing trajectories between samples in this latent space, or more simply to look at the distribution of some label in a manifold learned from the latent space. In such a way, it is possible to better understand which patterns were detected, or at which layer in the network classes begin to be separated (in the classification case). There is often no theoretical framework to illustrate these techniques, and then we referred to studies in the context of the medical application (*see Subheading 3.2* for references).

2.2.2 Input Optimization

Olah et al. [9] proposed to compute an input that maximizes the value of a feature map A (*see Fig. 5*). However, this technique leads to unrealistic images that may be themselves difficult to interpret, particularly for neuroimaging data. To have a better insight of the behavior of layers or filters, another simple technique illustrated by the same authors consists in isolating the inputs that led to the highest activation of A . The combination of both methods, displayed in Fig. 6, allows a better understanding of the concepts conveyed by A of a GoogLeNet trained on natural images.

2.3 Back-Propagation Methods

The goal of these interpretability methods is to link the value of an output node of interest o_c to the image X_0 given as input to a network. They do so by back-propagating a signal from o_c to X_0 :

Different optimization objectives show what different parts of a network are looking for.

n layer index
x, y spatial position
z channel index
k class index

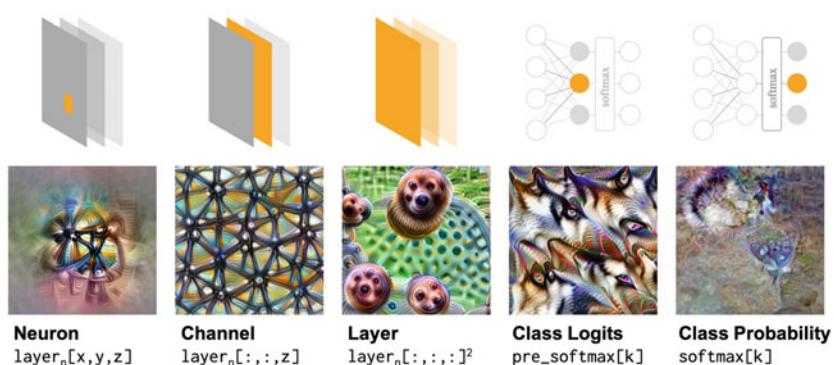


Fig. 5 Optimization of the input for different levels of feature maps. (Adapted from [9] (CC BY 4.0))

Dataset Examples show us what neurons respond to in practice



Optimization isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.



Baseball—or stripes?
mixed4a, Unit 6

Animal faces—or snouts?
mixed4a, Unit 240

Clouds—or fluffiness?
mixed4a, Unit 453

Buildings—or sky?
mixed4a, Unit 492

Fig. 6 Interpretation of a neuron of a feature map by optimizing the input associated with a bunch of training examples maximizing this neuron. (Adapted from [9] (CC BY 4.0))



Fig. 7 Attribution map of an image found with gradients back-propagation. (Adapted from [10]. Permission to reuse was kindly granted by the authors)

this process (backward pass) can be seen as the opposite operation than the one done when computing the output value from the input (forward pass).

Any property can be back-propagated as soon as its value at the level of a feature map $l - 1$ can be computed from its value in the feature map l . In this section, the back-propagated properties are gradients or the relevance of a node o_c .

2.3.1 Gradient Back-Propagation

During network training, gradients corresponding to each layer are computed according to the loss to update the weights. Then, we can see these gradients as the difference needed at the layer level to improve the final result: by adding this difference to the weights, the probability of the true class y increases.

In the same way, the gradients can be computed at the image level to find how the input should vary to change the value of o_c (*see* example on Fig. 7). This gradient computation was proposed by [10], in which the attribution map S_c corresponding to the input

image X_0 and the output node o_c is computed according to the following equation:

$$S_c = \frac{\partial o_c}{\partial X} \Big|_{X=X_0} \quad (1)$$

Due to its simplicity, this method is the most commonly used to interpret deep learning networks. Its attribution map is often called a “saliency map”; however, this term is also used in some articles to talk about any attribution map, and this is why we chose to avoid this term in this chapter.

This method was modified to derive many similar methods based on gradient computation described in the following paragraphs.

Gradient ⊙ Input This method is the point-wise product of the gradient map described at the beginning of the section and the input. Evaluated in [11], it was presented as an improvement of the gradients method, though the original paper does not give strong arguments on the nature of this improvement.

DeconvNet & Guided Back-Propagation The key difference between this procedure and the standard back-propagation method is the way the gradients are back-propagated through the ReLU layer.

The ReLU layer is a commonly used activation function that sets to 0 the negative input values and does not affect positive input values. The derivative of this function in layer l is the indicator function $\mathbb{1}_{A^{(l)} > 0}$: it outputs 1 (resp. 0) where the feature maps computed during the forward pass were positive (resp. negative).

Springenberg et al. [12] proposed to back-propagate the signal differently. Instead of applying the indicator function of the feature map $A^{(l)}$ computed during the forward pass, they directly applied ReLU to the back-propagated values $R^{(l+1)} = \frac{\partial o_c}{\partial A^{(l+1)}}$, which corresponds to multiplying it by the indicator function $\mathbb{1}_{R^{(l+1)} > 0}$. This “backward deconvnet” method allows back-propagating only the positive gradients, and, according to the authors, it results in a reconstructed image showing the part of the input image that is most strongly activating this neuron.

The guided back-propagation method (Eq. 4) combines the standard back-propagation (Eq. 2) with the backward deconvnet (Eq. 3): when back-propagating gradients through ReLU layers, a value is set to 0 if the corresponding top gradients or bottom data is negative. This adds an additional guidance to the standard back-propagation by preventing backward flow of negative gradients.

$$R^{(l)} = \mathbb{1}_{A^{(l)} > 0} * R^{(l+1)} \quad (2)$$

$$R^{(l)} = \mathbb{1}_{R^{(l+1)} > 0} * R^{(l+1)} \quad (3)$$

$$R^{(l)} = \mathbb{1}_{A^{(l)} > 0} * \mathbb{1}_{R^{(l+1)} > 0} * R^{(l+1)} \quad (4)$$

Any back-propagation procedure can be “guided,” as it only concerns the way ReLU functions are managed during back-propagation (this is the case, e.g., for guided Grad-CAM).

While it was initially adopted by the community, this method showed severe defects as discussed later in Subheading 4.

CAM & Grad-CAM In this setting, attribution maps are computed at the level of a feature map produced by a convolutional layer and then upsampled to be overlaid and compared with the input. The first method, class activation maps (CAM), was proposed by Zhou et al. [13] and can be only applied to CNNs with the following specific architecture:

1. A series of convolutions associated with activation functions and possibly pooling layers. These convolutions output a feature map A with N channels.
2. A global average pooling that extracts the mean value of each channel of the feature map produced by the convolutions.
3. A single fully connected layer

The CAM corresponding to o_c will be the mean of the channels of the feature map produced by the convolutions, weighted by the weights w_{kc} learned in the fully connected layer

$$S_c = \sum_{k=1}^N w_{kc} * A_k. \quad (5)$$

This map has the same size as A_k , which might be smaller than the input if the convolutional part performs downsampling operations (which is very often the case). Then, the map is upsampled to the size of the input to be overlaid on the input.

Selvaraju et al. [14] proposed an extension of CAM that can be applied to any architecture: Grad-CAM (illustrated on Fig. 8). As in CAM, the attribution map is a linear combination of the channels of a feature map computed by a convolutional layer. But, in this case, the weights of each channel are computed using gradient back-propagation

$$\alpha_{kc} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{\partial o_c}{\partial A_k(u)}. \quad (6)$$

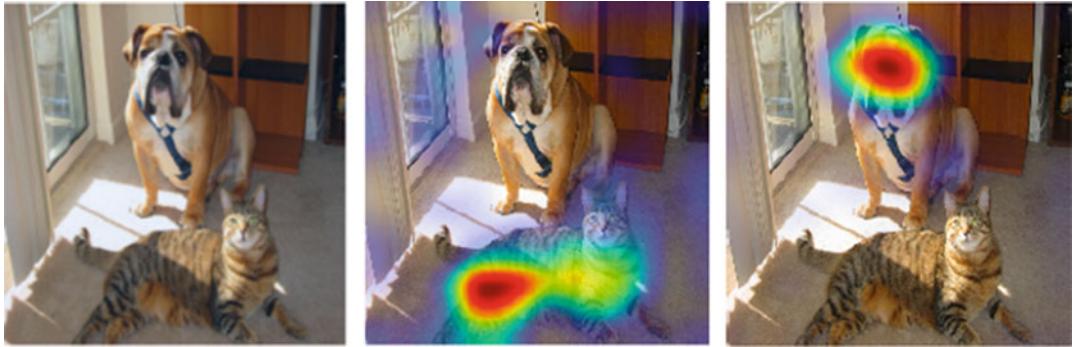


Fig. 8 Grad-CAM explanations highlighting two different objects in an image. **(a)** the original image, **(b)** the explanation based on the “dog” node, **(c)** the explanation based on the “cat” node. ©2017 IEEE. (Reprinted, with permission, from [14])

The final map is then the linear combination of the feature maps weighted by the coefficients. A ReLU activation is then applied to the result to only keep the features that have a positive influence on class c

$$S_c = \text{ReLU} \left(\sum_{k=1}^N \alpha_{kc} * A_k \right). \quad (7)$$

Similarly to CAM, this map is then upsampled to the input size.

Grad-CAM can be applied to any feature map produced by a convolution, but in practice the last convolutional layer is very often chosen. The authors argue that this layer is “the best compromise between high-level semantics and detailed spatial information” (the latter is lost in fully connected layers, as the feature maps are flattened).

Because of the upsampling step, CAM and Grad-CAM produce maps that are more human-friendly because they contain more connected zones, contrary to other attribution maps obtained with gradient back-propagation that can look very scattered. However, the smaller the feature maps A_k , the blurrier they are, leading to a possible loss of interpretability.

2.3.2 Relevance Back-Propagation

Instead of back-propagating gradients to the level of the input or of the last convolutional layer, Bach et al. [15] proposed to back-propagate the score obtained by a class c , which is called the relevance. This score corresponds to o_c after some post-processing (e.g., softmax), as its value must be positive if class c was identified in the input. At the end of the back-propagation process, the goal is to find the relevance R_u of each feature u of the input (e.g., of each pixel of an image) such that $o_c = \sum_{u \in U} R_u$.

In their paper, Bach et al. [15] take the example of a fully connected function defined by a matrix of weights w and a bias

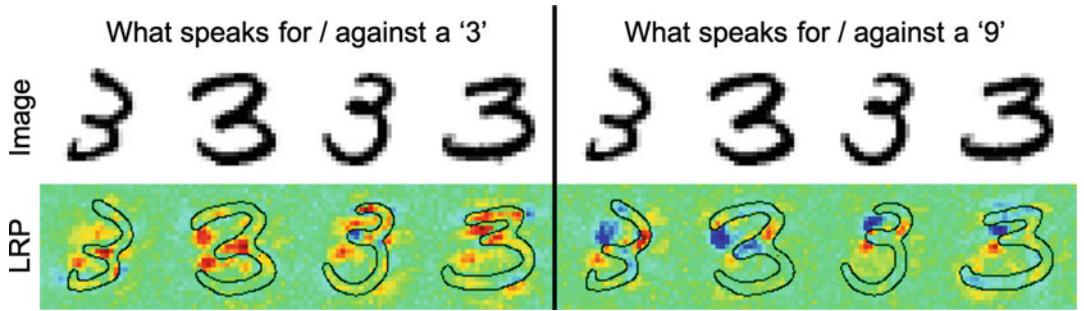


Fig. 9 LRP attribution maps explaining the decision of a neural network trained on MNIST. ©2017 IEEE. (Reprinted, with permission, from [16])

b at layer $l+1$. The value of a node v in feature map $A^{(l+1)}$ is computed during the forward pass by the given formula:

$$A^{(l+1)}(v) = b + \sum_{u \in \mathcal{U}} w_{uv} A^{(l)}(u) \quad (8)$$

During the back-propagation of the relevance, $R^{(l)}(u)$, the value of the relevance at the level of the layer $l+1$ is computed according to the values of the relevance $R^{(l+1)}(v)$ which are distributed according to the weights w learned during the forward pass and the values of $A^{(l)}(v)$:

$$R^{(l)}(u) = \sum_{v \in \mathcal{V}} R^{(l+1)}(v) \frac{A^{(l)}(u) w_{uv}}{\sum_{u' \in \mathcal{U}} A^{(l)}(u') w_{u'v}}. \quad (9)$$

The main issue of the method comes from the fact that the denominator may become (close to) zero, leading to the explosion of the relevance back-propagated. Moreover, it was shown by [11] that when all activations are piece-wise linear (such as ReLU or leaky ReLU), the layer-wise relevance (LRP) method reproduces the output of gradient \odot input, questioning the usefulness of the method.

This is why Samek et al. [16] proposed two variants of the standard LRP method [15]. Moreover they describe the behavior of the back-propagation in other layers than the linear ones (the convolutional one following the same formula as the linear). They illustrated their method with a neural network trained on MNIST (see Fig. 9). To simplify the equations in the following paragraphs, we now denote the weighted activations as $z_{uv} = A^{(l)}(u) w_{uv}$.

ϵ -rule The ϵ -rule integrates a parameter $\epsilon > 0$, used to avoid numerical instability. Though it avoids the case of a null denominator, this variant breaks the rule of relevance conservation across layers

$$R^{(l)}(u) = \sum_{v \in \mathcal{V}} R^{(l+1)}(v) \frac{z_{uv}}{\sum_{u' \in \mathcal{U}} z_{u'v} + \epsilon \times \text{sign}\left(\sum_{u' \in \mathcal{U}} z_{u'v}\right)}. \quad (10)$$

β -rule The β -rule keeps the conservation of the relevance by treating separately the positive weighted activations z_{uv}^+ from the negative ones z_{uv}^-

$$R^{(l)}(u) = \sum_{v \in \mathcal{V}} R^{(l+1)}(v) \left((1 + \beta) \frac{z_{uv}^+}{\sum_{u' \in \mathcal{U}} z_{u'v}^+} - \beta \frac{z_{uv}^-}{\sum_{u' \in \mathcal{U}} z_{u'v}^-} \right). \quad (11)$$

Though these two LRP variants improve the numerical stability of the procedure, they imply to choose the values of parameters that may change the patterns in the obtained attribution map.

Deep Taylor Decomposition Deep Taylor decomposition [17] was proposed by the same team as the one that proposed the original LRP method and its variants. It is based on similar principles as LRP: the value of the score obtained by a class c is back-propagated, but the back-propagation rule is based on first-order Taylor expansions.

The back-propagation from node v in at the level of $R^{(l+1)}$ to u at the level of $R^{(l)}$ can be written

$$R^{(l)}(u) = \sum_{v \in \mathcal{V}} \frac{\partial R^{(l+1)}(v)}{\partial A^{(l)}(u)} \Bigg|_{\tilde{A}_v^{(l)}(u)} \left(A^{(l)}(u) - \tilde{A}_v^{(l)}(u) \right). \quad (12)$$

This rule implies a root point $\tilde{A}_v^{(l)}(u)$ which is close to $A^{(l)}(u)$ and meets a set of constraints depending on v .

2.4 Perturbation Methods

Instead of relying on a backward pass (from the output to the input) as in the previous section, perturbation methods rely on the difference between the value of o_c computed with the original inputs and a locally perturbed input. This process is less abstract for humans than back-propagation methods as we can reproduce it ourselves: if the part of the image that is needed to find the good output is hidden, we are also not able to predict correctly. Moreover, it is model-agnostic and can be applied to any algorithm or deep learning architecture.

The main drawback of these techniques is that the nature of the perturbation is crucial, leading to different attribution maps depending on the perturbation function used. Moreover, Montavon et al. [18] suggest that the perturbation rule should

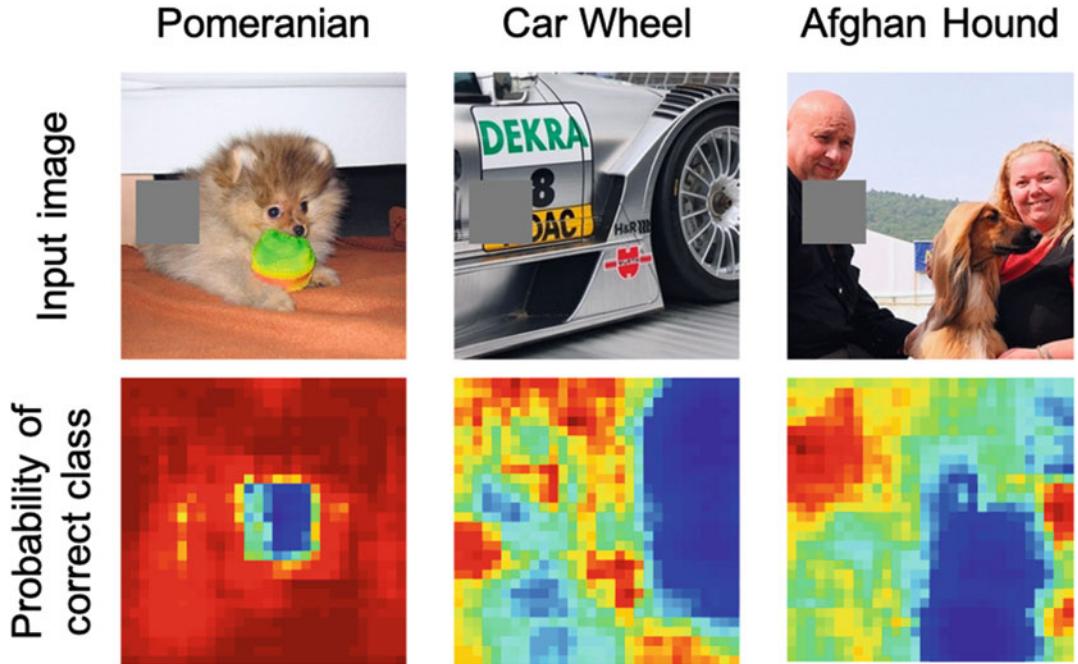


Fig. 10 Attribution maps obtained with standard perturbation. Here the perturbation is a gray patch covering a specific zone of the input as shown in the left column. The attribution maps (second row) display the probability of the true label: the lower the value, the most important it is for the network to correctly identify the label. This kind of perturbation takes the perturbed input out of the training distribution. (Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, ECCV 2014: Visualizing and Understanding Convolutional Networks, [19], 2014)

keep the perturbed input in the training data distribution. Indeed, if it is not the case, one cannot know if the network performance dropped because of the location or the nature of the perturbation.

2.4.1 Standard Perturbation

Zeiler and Fergus [19] proposed the most intuitive method relying on perturbations. This standard perturbation procedure consists in removing information locally in a specific zone of an input X_0 and evaluating if it modifies the output node o_v . The more the perturbation degrades the task performance, the more crucial this zone is for the network to correctly perform the task. To obtain the final attribution map, the input is perturbed according to all possible locations. Examples of attribution maps obtained with this method are displayed in Fig. 10.

As evaluating the impact of the perturbation at each pixel location is computationally expensive, one can choose not to perturb the image at each pixel location but to skip some of them (i.e., scan the image with a stride > 1). This will lead to a smaller attribution map, which needs to be upsampled to be compared to the original input (in the same way as CAM & Grad-CAM).

However, in addition to the problem of the nature of the perturbation previously mentioned, this method presents two drawbacks:

- The attribution maps depend on the size of the perturbation: if the perturbation becomes too large, the perturbation is not local anymore; if it is too small, it is not meaningful anymore (a pixel perturbation cannot cover a pattern).
- Input pixels are considered independently from each other: if the result of a network relies on a combination of pixels that cannot all be covered at the same time by the perturbation, their influence may not be detected.

2.4.2 Optimized Perturbation

To deal with these two issues, Fong and Vedaldi [2] proposed to optimize a perturbation mask covering the whole input. This perturbation mask m has the same size as the input X_0 . Its application is associated with a perturbation function Φ and leads to the computation of the perturbed input X_0^m . Its value at a coordinate u reflects the quantity of information remaining in the perturbed image:

- If $m(u) = 1$, the pixel at location u is not perturbed and has the same value in the perturbed input as in the original input ($X_0^m(u) = X_0(u)$).
- If $m(u) = 0$, the pixel at location u is fully perturbed and the value in the perturbed image is the one given by the perturbation function only ($X_0^m(u) = \Phi(X_0)(u)$).

This principle can be extended to any value between 0 and 1 with the a linear interpolation

$$X_0^m(u) = m(u)X_0(u) + (1 - m(u))\Phi(X_0)(u). \quad (13)$$

Then, the goal is to optimize this mask m according to three criteria:

1. The perturbed input X_0^m should lead to the lowest performance possible.
2. The mask m should perturb the minimum number of pixels possible.
3. The mask m should produce connected zones (i.e., avoid the scattered aspect of gradient maps).

These three criteria are optimized using the following loss:

$$f(X_0^m) + \lambda_1 \|1 - m\|_{\beta_1}^{\beta_1} + \lambda_2 \|\nabla m\|_{\beta_2}^{\beta_2} \quad (14)$$

with f a function that decreases as the performance of the network decreases.

However, the method also presents two drawbacks:



Fig. 11 In this example, the network learned to classify objects in natural images. Instead of masking the maypole at the center of the image, it creates artifacts in the sky to degrade the performance of the network. ©2017 IEEE. (Reprinted, with permission, from [2])

- The values of hyperparameters must be chosen ($\lambda_1, \lambda_2, \beta_1, \beta_2$) to find a balance between the three optimization criteria of the mask.
- The mask may not highlight the most important pixels of the input but instead create artifacts in the perturbed image to artificially degrade the performance of the network (see Fig. 11).

2.5 Distillation

Approaches described in this section aim at developing a transparent method to reproduce the behavior of a black box one. Then it is possible to consider simple interpretability methods (such as weight visualization) on the transparent method instead of considering the black box.

2.5.1 Local Approximation

LIME Ribeiro et al. [1] proposed local interpretable model-agnostic explanations (LIME). This approach is:

- **Local**, as the explanation is valid in the vicinity of a specific input X_0
- **Interpretable**, as an interpretable model g (linear model, decision tree...) is computed to reproduce the behavior of f on X_0
- **Model-agnostic**, as it does not depend on the algorithm trained

This last property comes from the fact that the vicinity of X_0 is explored by sampling variations of X_0 that are perturbed versions of X_0 . Then LIME shares the advantage (model-agnostic) and drawback (perturbation function dependent) of perturbation methods presented in Subheading 2.4. Moreover, the authors specify that, in the case of images, they group pixels of the input in d super-pixels (contiguous patches of similar pixels).

The loss to be minimized to find g specific to the input X_0 is the following:

$$\mathcal{L}(f, g, \pi_{X_0}) + \Omega(g), \quad (15)$$

where π_{X_0} is a function that defines the locality of X_0 (i.e., $\pi_{X_0}(X)$ decreases as X becomes closer to X_0), \mathcal{L} measures how unfaithful \mathcal{g} is in approximating f according to π_{X_0} , and Ω is a measure of the complexity of \mathcal{g} .

Ribeiro et al. [1] limited their search to sparse linear models; however, other assumptions could be made on \mathcal{g} .

\mathcal{g} is not applied to the input directly but to a binary mask $m \in \{0, 1\}^d$ that transforms the input X in X^m and is applied according to a set of d super-pixels. For each super-pixel u :

1. If $m(u) = 1$, the super-pixel u is not perturbed.
2. If $m(u) = 0$, the super-pixel u is perturbed (i.e., it is grayed).

They used

$$\pi_{X_0}(X) = \exp \frac{(X - X_0)^2}{\sigma^2}$$

and

$$\mathcal{L}(f, \mathcal{g}, \pi_{X_0}) = \sum_m \pi_{X_0}(X_0^m) * (f(X_0^m) - \mathcal{g}(m))^2.$$

Finally $\Omega(\mathcal{g})$ is the number of non-zero weights of \mathcal{g} , and its value is limited to K . This way they select the K super-pixels in X_0 that best explain the algorithm result $f(X_0)$.

SHAP Lundberg and Lee [20] proposed SHAP (SHapley Additive exPlanations), a theoretical framework that encompasses several existing interpretability methods, including LIME. In this framework each of the N features (again, super-pixels for images) is associated with a coefficient ϕ that denotes its contribution to the result. The contribution of each feature is evaluated by perturbing the input X_0 with a binary mask m (see paragraph on LIME). Then the goal is to find an interpretable model \mathcal{g} specific to X_0 , such that

$$\mathcal{g}(m) = \phi_0 + \sum_i^N \phi_i m_i \quad (16)$$

with ϕ_0 being the output when the input is fully perturbed.

The authors look for an expression of ϕ that respects three properties:

- **Local accuracy** \mathcal{g} and f should match in the vicinity of X_0 : $\mathcal{g}(m) = f(X_0^m)$.
- **Missingness** Perturbed features should not contribute to the result: $m_i = 0 \rightarrow \phi_i = 0$.
- **Consistency** Let's denote as $m \setminus i$ the mask m in which $m_i = 0$. For any two models f^1 and f^2 , if

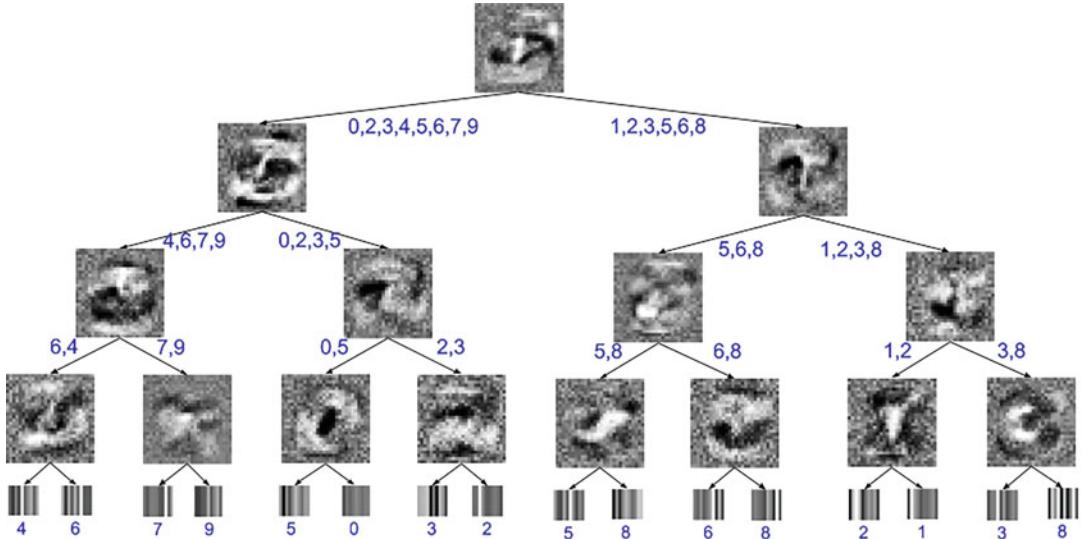


Fig. 12 Visualization of a soft decision tree trained on MNIST. (Adapted from [21]. Permission to reuse was kindly granted by the authors)

$$f^1(X_0^m) - f^1(X_0^{m \setminus i}) \geq f^2(X_0^m) - f^2(X_0^{m \setminus i})$$

then for all $m \in \{0, 1\}^N$ $\phi_i^1 \geq \phi_i^2$ (ϕ^k are the coefficients associated with model f^k).

Lundberg and Lee [20] show that only one expression is possible for the coefficients ϕ , which can be approximated with different algorithms:

$$\phi_i = \sum_{m \in \{0, 1\}^N} \frac{|m|!(N - |m| - 1)!}{N!} [f(X_0^m) - f(X_0^{m \setminus i})]. \quad (17)$$

2.5.2 Model Translation

Contrary to local approximation, which provides an explanation according to a specific input X_0 , model translation consists in finding a transparent model that reproduces the behavior of the black box model on the whole data set.

As it was rarely employed in neuroimaging frameworks, this section only discusses the distillation to decision trees proposed in [21] (preprint). For a more extensive review of model translation methods, we refer the reader to [5].

After training a machine learning system f , a binary decision tree g is trained to reproduce its behavior. This tree is trained on a set of inputs X , and each inner node i learns a matrix of weights w_i and biases b_i . The forward pass of X in the node i of the tree is as follows: if $\text{sigmoid}(w_i X + b_i) > 0.5$, then the right leaf node is chosen, else the left leaf node is chosen. After the end of the decision tree's training, it is possible to visualize at which level which classes were separated to better understand which classes are similar for the

network. It is also possible to visualize the matrices of weights learned by each inner node to identify patterns learned at each class separation. An illustration of this distillation process, on the MNIST data set (hand-written digits), can be found in Fig. 12.

2.6 Intrinsic

Contrary to the previous sections in which interpretability methods could be applied to (almost) any network after the end of the training procedure, the following methods require to design the framework before the training phase, as the interpretability components and the network are trained simultaneously. In the papers presented in this Subheading [22–24], the advantages of these methods are dual: they improve both the interpretability and performance of the network. However, the drawback is that they have to be implemented before training the network, and then they cannot be applied in all cases.

2.6.1 Attention Modules

Attention is a concept in machine learning that consists in producing an attribution map from a feature map and using it to improve learning of another task (such as classification, regression, reconstruction...) by making the algorithm focus on the part of the feature map highlighted by the attribution map.

In the deep learning domain, we take as reference [22], in which a network is trained to produce a descriptive caption of natural images. This network is composed of three parts:

1. A convolutional encoder that reduces the dimension of the input image to the size of the feature maps A

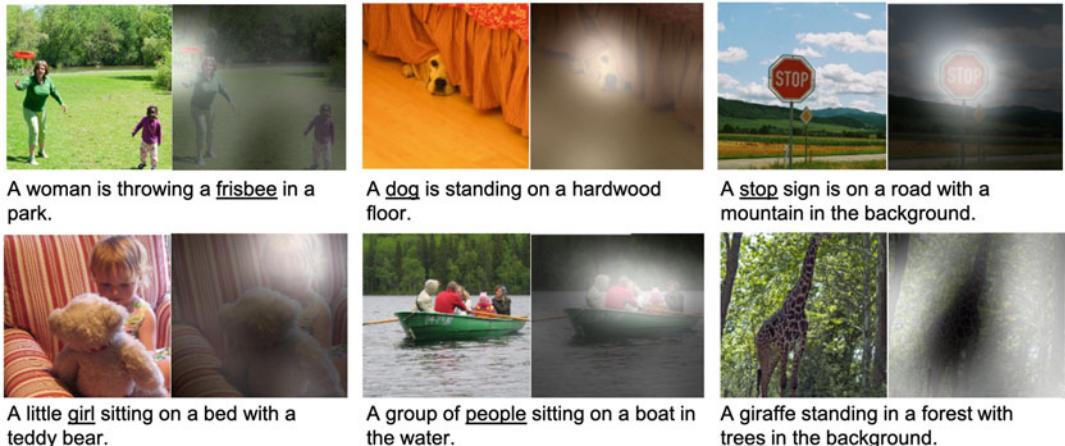


Fig. 13 Examples of images correctly captioned by the network. The focus of the attribution map is highlighted in white and the associated word in the caption is underlined. (Adapted from [22]. Permission to reuse was kindly granted by the authors)

2. An attention module that generates an attribution map S_t from A and the previous hidden state of the long short-term memory (LSTM) network
3. An LSTM decoder that computes the caption from its previous hidden state, the previous word generated, A and S_t

As S_t is of the same size as A (smaller than the input), the result is then upsampled to be overlaid on the input image. As one attribution map is generated per word generated by the LSTM, it is possible to know where the network focused when generating each word of the caption (*see* Fig. 13). In this example, the attribution map is given to a LSTM, which uses it to generate a context vector z_t by applying a function ϕ to A and S_t .

More generally in CNNs, the point-wise product of the attribution map S and the feature map A is used to generate the refined feature map A' which is given to the next layers of the network. Adding an attention module implies to make new choices for the architecture of the model: its location (on lower or higher feature maps) may impact the performance of the network. Moreover, it is possible to stack several attention modules along the network, as it was done in [23].

2.6.2 Modular Transparency

Contrary to the studies of the previous sections, the frameworks of these categories are composed of several networks (modules) that interact with each other. Each module is a black box, but the transparency of the function, or the nature of the interaction

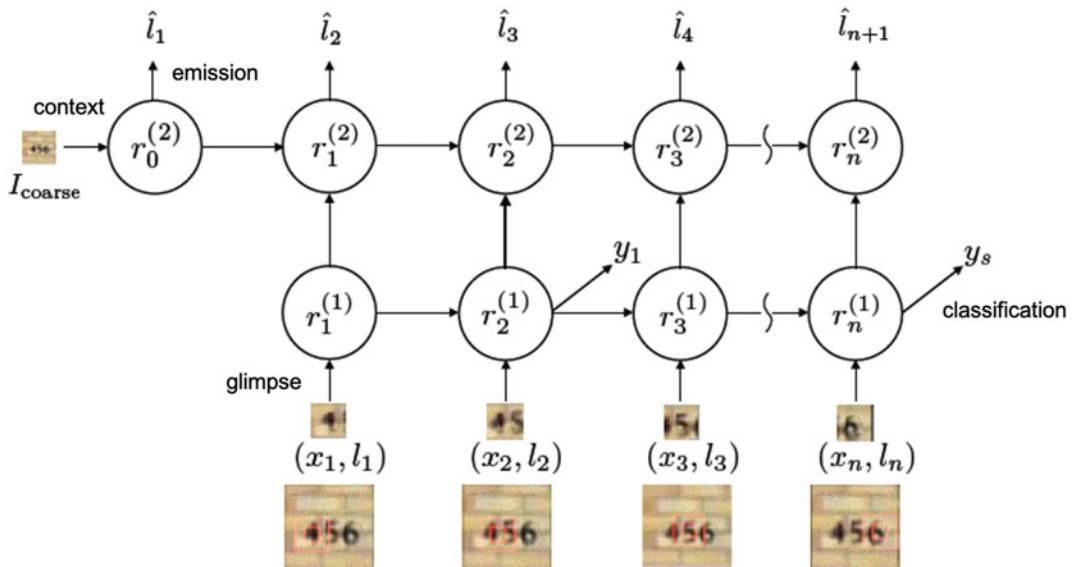


Fig. 14 Framework with modular transparency browsing an image to compute the output at the global scale. (Adapted from [24]. Permission to reuse was kindly granted by the authors)

between them, allows understanding how the system works globally and extracting interpretability metrics from it.

A large variety of setups can be designed following this principle, and it is not possible to draw a more detailed general rule for this section. We will take the example described in [24], which was adapted to neuroimaging data (*see* Subheading 3.6), to illustrate this section, though it may not be representative of all the aspects of modular transparency.

Ba et al. [24] proposed a framework (illustrated in Fig. 14) to perform the analysis of an image in the same way as a human, by looking at successive relevant locations in the image. To perform this task, they assemble a set of networks that interact together:

- **Glimpse network** This network takes as input a patch of the input image and the location of its center to output a context vector that will be processed by the recurrent network. Then this vector conveys information on the main features in a patch and its location.
- **Recurrent network** This network takes as input the successive context vectors and update its hidden state that will be used to find the next location to look at and to perform the learned task at the global scale (in the original paper a classification of the whole input image).
- **Emission network** This network takes as input the current state of the recurrent network and outputs the next location to look at. This will allow computing the patch that will feed the glimpse network.
- **Context network** This network takes as input the whole input at the beginning of the task and outputs the first context vector to initialize the recurrent network.
- **Classification network** This network takes as input the current state of the recurrent network and outputs a prediction for the class label.

The global framework can be seen as interpretable as it is possible to review the successive processed locations.

2.7 Interpretability Metrics

To evaluate the reliability of the methods presented in the previous sections, one cannot only rely on qualitative evaluation. This is why interpretability metrics that evaluate attribution maps were proposed. These metrics may evaluate different properties of attribution maps.

- **Fidelity** evaluates if the zones highlighted by the map influence the decision of the network.
- **Sensitivity** evaluates how the attribution map changes according to small changes in the input X_0 .

- **Continuity** evaluates if two close data points lead to similar attribution maps.

In the following, Γ is an interpretability method computing an attribution map S of the black box network f and an input X_0 .

2.7.1 (In)fidelity

Yeh et al. [25] proposed a measure of infidelity of Γ based on perturbations applied according to a vector m of the same shape as the attribution map S . The explanation is infidel if perturbations applied in zones highlighted by S on X_0 lead to negligible changes in $f(X_0^m)$ or, on the contrary, if perturbations applied in zones not highlighted by S on X_0 lead to significant changes in $f(X_0^m)$. The associated formula is

$$\text{INFD}(\Gamma, f, X_0) = \mathbb{E}_m \left[\sum_i \sum_j m_{ij} \Gamma(f, X_0)_{ij} - (f(X_0) - f(X_0^m))^2 \right]. \quad (18)$$

2.7.2 Sensitivity

Yeh et al. [25] also gave a measure of sensitivity. As suggested by the definition, it relies on the construction of attribution maps according to inputs similar to X_0 : \tilde{X}_0 . As changes are small, sensitivity depends on a scalar ϵ set by the user, which corresponds to the maximum difference allowed between X_0 and \tilde{X}_0 . Then sensitivity corresponds to the following formula:

$$\text{SENS}_{\max}(\Gamma, f, X_0, \epsilon) = \max_{\|\tilde{X}_0 - X_0\| \leq \epsilon} \|\Gamma(f, \tilde{X}_0) - \Gamma(f, X_0)\|. \quad (19)$$

2.7.3 Continuity

Continuity is very similar to sensitivity, except that it compares different data points belonging to the input domain \mathcal{X} , whereas sensitivity may generate similar inputs with a perturbation method. This measure was introduced in [18] and can be computed using the following formula:

$$\text{CONT}(\Gamma, f, \mathcal{X}) = \max_{X_1, X_2 \in \mathcal{X} \& X_1 \neq X_2} \frac{\|\Gamma(f, X_1) - \Gamma(f, X_2)\|_1}{\|X_1 - X_2\|_2}. \quad (20)$$

As these metrics rely on perturbation, they are also influenced by the nature of the perturbation and may lead to different results, which is a major issue (*see* Subheading 4). Other metrics were also proposed and depend on the task learned by the network: for example, in the case of a classification, statistical tests can be conducted between attribution maps of different classes to assess whether they differ according to the class they explain.

3 Application of Interpretability Methods to Neuroimaging Data

In this section, we provide a non-exhaustive review of applications of interpretability methods to neuroimaging data. In most cases, the focus of articles is prediction/classification rather than the interpretability method, which is just seen as a tool to analyze the results. Thus, authors do not usually motivate their choice of an interpretability method. Another key consideration here is the spatial registration of brain images, which enables having brain regions roughly at the same position between subjects. This technique is of paramount importance as attribution maps computed for registered images can then be averaged or used to automatically determine the most important brain areas, which would not be possible with unaligned images. All the studies presented in this section are summarized in Table 3.

Table 3
Summary of the studies applying interpretability methods to neuroimaging data which are presented in Subheading 3

Study	Data set	Modality	Task	Interpretability method	Section
Abrol et al. [28]	ADNI	T1w	AD classification	FM visualization, Perturbation	3.2, 3.4
Bae et al. [32]	ADNI	sMRI	AD classification	Perturbation	3.4
Ball et al. [33]	PING	T1w	Age prediction	Weight visualization, SHAP	3.1, 3.5
Biffi et al. [29]	ADNI	T1w	AD classification	FM visualization	3.2
Böhle et al. [34]	ADNI	T1w	AD classification	LRP, Guided back-propagation	3.3
Burduja et al. [35]	RSNA	CT scan	Intracranial Hemorrhage detection	Grad-CAM	3.3
Cecotti and Gräser [26]	in-house	EEG	P300 signals detection	Weight visualization	3.1
Dyrba et al. [36]	ADNI	T1w	AD classification	DeconvNet, Deep Taylor decomposition, Gradient ⊖ Input, LRP, Grad-CAM	3.3
Eitel and Ritter [37]	ADNI	T1w	AD classification	Gradient ⊖ Input, Guided back-propagation, LRP, Perturbation	3.3, 3.4

(continued)

Table 3
(continued)

Study	Data set	Modality	Task	Interpretability method	Section
Eitel et al. [38]	ADNI, in-house	T1w	Multiple Sclerosis detection	Gradient ⊕ Input, LRP	3.3
Fu et al. [39]	CQ500, RSNA	CT scan	Detection of Critical Findings in Head CT scan	Attention mechanism	3.6
Gutiérrez-Becker and Wachinger [40]	ADNI	T1w	AD classification	Perturbation	3.4
Hu et al. [41]	ADNI, NIFD	T1w	AD/CN/FTD classification	Guided back- propagation	3.3
Jin et al. [42]	ADNI, in-house	T1w	AD classification	Attention mechanism	3.6
Lee et al. [43]	ADNI	T1w	AD classification	Modular transparency	3.6
Leming et al. [31]	OpenFMRI, ADNI, ABIDE, ABIDE II, ABCD, NDAR ICBM, UK Biobank, 1000FC	fMRI	Autism classification Sex classification Task vs rest classification	FM visualization, Grad-CAM	3.2, 3.3
Magesh et al. [44]	PPMI	SPECT	Parkinson's disease detection	LIME	3.5
Martinez-Murcia et al. [30]	ADNI	T1w	AD classification Prediction of neuropsychological tests & other clinical variables	FM visualization	3.2
Nigri et al. [45]	ADNI, AIBL	T1w	AD classification	Perturbation, Swap test	3.4
Oh et al. [27]	ADNI	T1w	AD classification	FM visualization, Standard back- propagation, Perturbation	3.2, 3.3, 3.4
Qiu et al. [46]	ADNI, AIBL, FHS, NACC	T1w	AD classification	Modular transparency	3.6
Ravi et al. [47]	ADNI	T1w	CN/MCI/AD reconstruction	Modular transparency	3.6

(continued)

Table 3
(continued)

Study	Data set	Modality	Task	Interpretability method	Section
Rieke et al. [48]	ADNI	T1w	AD classification	Standard back-propagation, Guided back-propagation, Perturbation, Brain area occlusion	3.3, 3.4
Tang et al. [49]	UCD-ADC, Brain Bank	Histology	Detection of amyloid- β pathology	Guided back-propagation, Perturbation	3.3, 3.4
Wood et al. [50]	ADNI	T1w	AD classification	Modular transparency	3.6

Data sets: 1000FC, 1000 Functional Connectomes; ABCD, Adolescent Brain Cognitive Development; ABIDE, Autism Brain Imaging Data Exchange; ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarkers and Lifestyle; FHS, Framingham Heart Study; ICBM, International Consortium for Brain Mapping; NACC, National Alzheimer's Coordinating Center; NDAR, National Database for Autism Research; NIFD, frontotemporal lobar degeneration neuroimaging initiative; PING, Pediatric Imaging, Neurocognition and Genetics; PPMI, Parkinson's Progression Markers Initiative; RSNA, Radiological Society of North America 2019 Brain CT Hemorrhage data set; UCD-ADC Brain Bank, University of California Davis Alzheimer's Disease Center Brain Bank

Modalities: CT, computed tomography; EEG, electroencephalography; fMRI, functional magnetic resonance imaging; sMRI, structural magnetic resonance imaging; SPECT, single-photon emission computed tomography; T1w, T1-weighted [magnetic resonance imaging]

Tasks: AD, Alzheimer's disease; CN, cognitively normal; FTD, frontotemporal dementia; MCI, mild cognitive impairment

Interpretability methods: FM, feature maps; Grad-CAM, gradient-weighted class activation mapping; LIME, local interpretable model-agnostic explanations; LRP, layer-wise relevance; SHAP, SHapley Additive exPlanations

This section ends with the presentation of benchmarks conducted in the literature to compare different interpretability methods in the context of brain disorders.

3.1 Weight Visualization Applied to Neuroimaging

As the focus of this chapter is on non-transparent models, such as deep learning ones, weight visualization was only rarely found. However, this was the method chosen by Cecotti and Gräser [26], who developed a CNN architecture adapted to weight visualization to detect P300 signals in electroencephalograms (EEG). The input of this network is a matrix with rows corresponding to the 64 electrodes and columns to 78 time points. The two first layers of the networks are convolutions with rectangular filters: the first filters (size 1×64) combine the electrodes, whereas the second ones (13×1) find time patterns. Then, it is possible to retrieve a coefficient per electrode by summing the weights associated with this electrode across the different filters and to visualize the results in the electroencephalogram space as shown in Fig. 15.

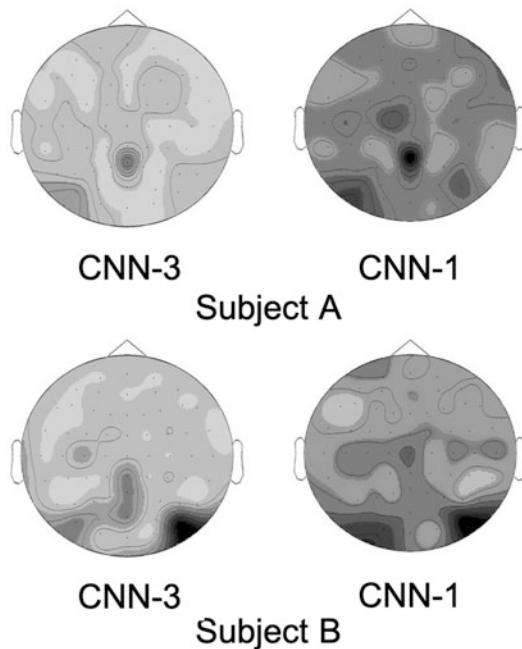


Fig. 15 Relative importance of the electrodes for signal detection in EEG using two different architectures (CNN-1 and CNN-3) and two subjects (A and B) using CNN weight visualization. Dark values correspond to weights with a high absolute value while white values correspond to weights close to 0. ©2011 IEEE. (Reprinted, with permission, from [26])

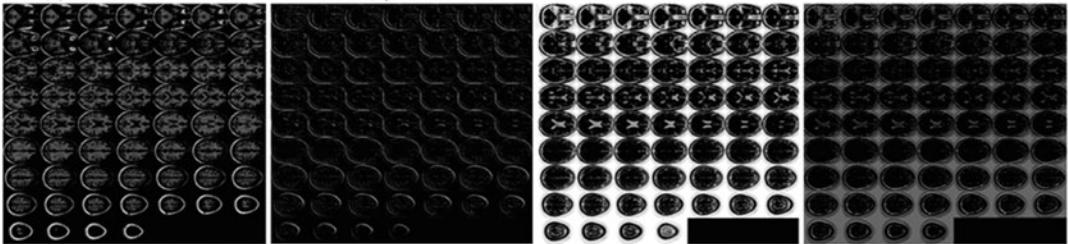
3.2 Feature Map Visualization Applied to Neuroimaging

Contrary to the limited application of weight visualization, there is an extensive literature about leveraging individual feature maps and latent spaces to better understand how models work. This goes from the visualization of these maps or their projections [27–29], to the analysis of neuron behavior [30, 31], through sampling in latent spaces [29].

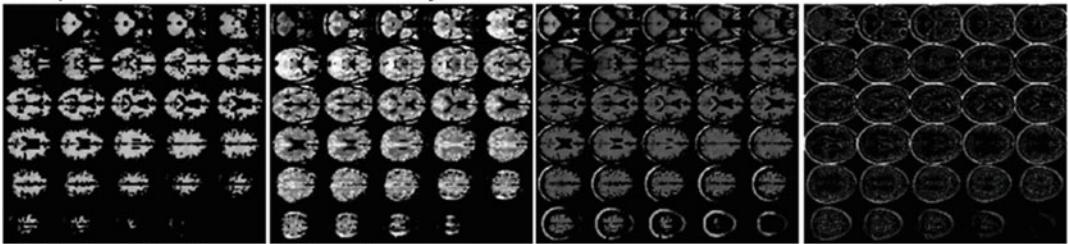
Oh et al. [27] displayed the feature maps associated with the convolutional layers of CNNs trained for various Alzheimer's disease status classification tasks (Fig. 16). In the first two layers, the extracted features were similar to white matter, cerebrospinal fluid, and skull segmentations, while the last layer showcased sparse, global, and nearly binary patterns. They used this example to emphasize the advantage of using CNNs to extract very abstract and complex features rather than using custom algorithms for feature extraction [27].

Another way to visualize a feature map is to project it in a two- or three-dimensional space to understand how it is positioned with respect to other feature maps. Abrol et al. [28] projected the features obtained after the first dense layer of a ResNet architecture onto a two-dimensional space using the classical t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction technique. For the classification task of Alzheimer's disease

Outputs of first convolutional layers



Outputs of second convolutional layers



Outputs of third convolutional layers

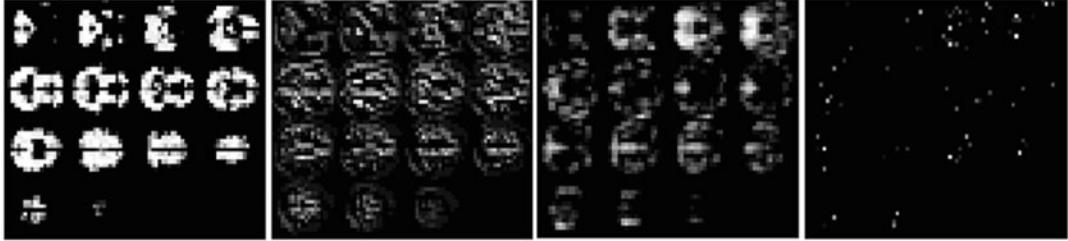


Fig. 16 Representation of a selection of feature maps (outputs of 4 filters on 10 for each layer) obtained for a single individual. (Adapted from [27] (CC BY 4.0))

statuses, they observed that the projections were correctly ordered according to the disease severity, supporting the correctness of the model [28]. They partitioned these projections into three groups: Far-AD (more extreme Alzheimer's Disease patients), Far-CN (more extreme Cognitively Normal participants), and Fused (a set of images at the intersection of AD and CN groups). Using a t-test, they were able to detect and highlight voxels presenting significant differences between groups (Fig. 17).

Biffi et al. [29] not only used feature map visualization but also sampled the feature space. Indeed, they trained a ladder variational autoencoder framework to learn hierarchical latent representations of 3D hippocampal segmentations of control subjects and Alzheimer's disease patients. A multilayer perceptron was jointly trained on top of the highest two-dimensional latent space to classify anatomical shapes. While lower spaces needed a dimensionality reduction technique (i.e., t-SNE), the highest latent space could directly be visualized, as well as the anatomical variability it captured in the initial input space, by leveraging the generative process of the model. This sampling enabled an easy visualization and quantification of the anatomical differences between each class.

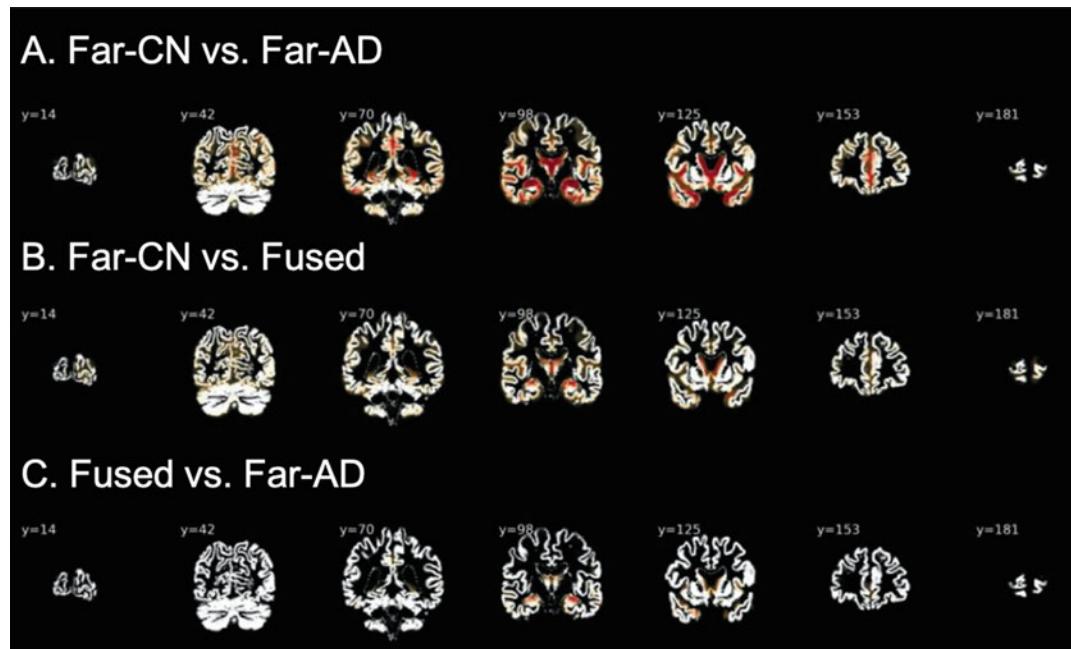


Fig. 17 Difference in neuroimaging space between groups defined thanks to t-SNE projection. Voxels showing significant differences post false discovery rate (FDR) correction ($p < 0.05$) are highlighted. (Reprinted from Journal of Neuroscience Methods, 339, [28], 2020, with permission from Elsevier)

Finally, it may be very informative to better understand the behavior of neurons and what they are encoding. After training deep convolutional autoencoders to reconstruct MR images, segmented gray matter maps, and white matter maps, Martinez-Murcia et al. [30] computed correlations between each individual hidden neuron value and clinical information (e.g., age, mini-mental state examination) which allowed them to determine to which extent this information was encoded in the latent space. This way they determined which clinical data was the most strongly associated. Using a collection of nine different MRI data sets, Leming et al. [31] trained CNNs for various classification tasks (autism vs typically developing, male vs female, and task vs rest). They computed a diversity coefficient for each filter of the second layer based on its output feature map. They counted how many different data sets maximally activated each value of this feature map: if they were mainly activated by one source of data, the coefficient would be close to 0, whereas if they were activated by all data sets, it would be close to 1. This allows assessing the layer stratification, i.e., to understand if a given filter was mostly maximally activated by one phenotype or by a diverse population. They found out that a few filters were only maximally activated by images from a single MRI data set and that the diversity coefficient was not normally distributed across filters, having generally two peaks at the

beginning and at the end of the spectrum, respectively, exhibiting the stratification and strongly diverse distribution of the filters.

3.3 Back-Propagation Methods Applied to Neuroimaging

3.3.1 Single Interpretation

Back-propagation methods are the most popular methods to interpret models, and a wide range of these algorithms have been used to study brain disorders: standard and guided back-propagation [27, 34, 37, 41, 48], gradientinput [36–38], Grad-CAM [35, 36], guided Grad-CAM [49], LRP [34, 36–38], DeconvNet [36], and deep Taylor Decomposition [36].

Some studies implemented a single back-propagation method and exploited it to find which brain regions are exploited by their algorithm [27, 31, 41], to validate interpretability methods [38], or to provide attribution maps to physicians to improve clinical guidance [35].

Oh et al. [27] used the standard back-propagation method to interpret CNNs for classification of Alzheimer's disease statuses. They showed that the attribution maps associated with the prediction of the conversion of prodromal patients to dementia included more complex representations, less focused on the hippocampi, than the ones associated with classification between demented patients from cognitively normal participants (*see* Fig. 18). In the context of autism, Leming et al. [31] used the Grad-CAM

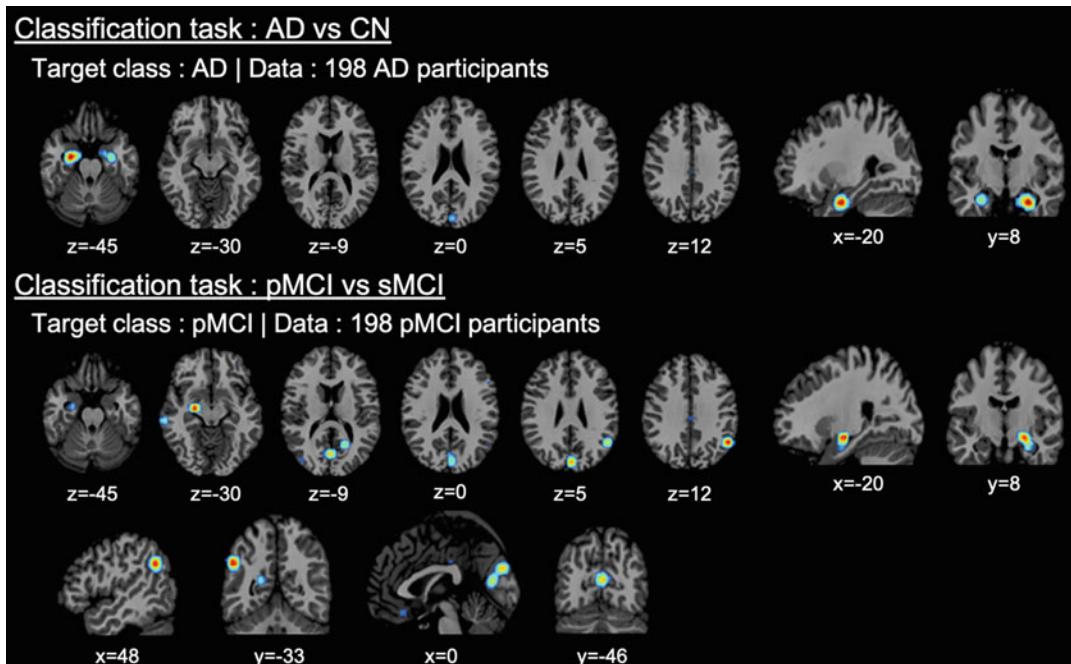


Fig. 18 Distribution of discriminant regions obtained with gradient back-propagation in the classification of demented patients and cognitively normal participants (top part, AD vs CN) and the classification of stable and progressive mild cognitive impairment (bottom part, sMCI vs pMCI). (Adapted from [27] (CC BY 4.0))

algorithm to determine the most important brain connections from functional connectivity matrices. However, the authors pointed out that without further work, this visualization method did not allow understanding the underlying reason of the attribution of a given feature: for instance, one cannot know if a set of edges is important because it is under-connected or over-connected. Finally, Hu et al. [41] used attribution maps produced by guided back-propagation to quantify the difference in the regions used by their network to characterize Alzheimer's disease or frontotemporal dementia.

The goal of Eitel et al. [38] was different. Instead of identifying brain regions related to the classification task, they exhibited with LRP that transfer learning between networks trained on different diseases (Alzheimer's disease to multiple sclerosis) and different MRI sequences enabled obtaining attribution maps focused on a smaller number of lesion areas. However, the authors pointed out that it would be necessary to confirm their results on larger data sets.

Finally, Burduja et al. [35] trained a CNN-LSTM model to detect various hemorrhages from brain computed tomography (CT) scans. For each positive slice coming from controversial or difficult scans, they generated Grad-CAM-based attribution maps and asked a group of radiologists to classify them as correct, partially correct, or incorrect. This classification allowed them to determine patterns for each class of maps and better understand which characteristics radiologists expected from these maps to be considered as correct and thus useful in practice. In particular, radiologists described maps including any type of hemorrhage as incorrect as soon as some of the hemorrhages were not highlighted, while the model only needed to detect one hemorrhage to correctly classify the slice as pathological.

3.3.2 Comparison of Several Interpretability Methods

Papers described in this section used several interpretability methods and compared them in their particular context. However, as the benchmark of interpretability methods is the focus of Subheading 4.3, which also include other types of interpretability than back-propagation, we will only focus here on what conclusions were drawn from the attribution maps.

Dyrba et al. [36] compared DeconvNet, guided back-propagation, deep Taylor decomposition, gradient \odot input, LRP (with various rules), and Grad-CAM methods for classification of Alzheimer's disease, mild cognitive impairment, and normal cognition statuses. In accordance with the literature, they obtained a highest attention given to the hippocampus for both prodromal and demented patients.

Böhle et al. [34] compared two methods, LRP with β -rule and guided back-propagation for Alzheimer's disease status classification. They found that LRP attribution maps highlight the individual differences between patients and then that they could be used as a tool for clinical guidance.

3.4 Perturbation Methods Applied to Neuroimaging

3.4.1 Variants of the Perturbation Method Tailored to Neuroimaging

The standard perturbation method has been widely used in the study of Alzheimer's disease [32, 37, 45, 48] and related symptoms (amyloid- β pathology) [49]. However, most of the time, authors do not train their model with perturbed images. Hence, to generate explanation maps, the perturbation method uses images outside the distribution of the training set, which may call into question the relevance of the predictions and thus the reliability of attention maps.

Several variations of the perturbation method have been developed to adapt to neuroimaging data. The most common variation in brain imaging is the brain area perturbation method, which consists in perturbing entire brain regions according to a given brain atlas, as done in [27, 28, 48]. In their study of Alzheimer's disease, Abrol et al. [28] obtained high values in their attribution maps for the usually discriminant brain regions, such as the hippocampus, the amygdala, the inferior and superior temporal gyri, and the fusiform gyrus. Rieke et al. [48] also obtained results in accordance with the medical literature and noted that the brain area perturbation method led to a less scattered attribution map than the standard method (Fig. 19). Oh et al. [27] used the method to compare the attribution maps of two different tasks: (1) demented patients vs cognitively normal participants and (2) stable vs progressive mild cognitively impaired patients and noted that the regions targeted

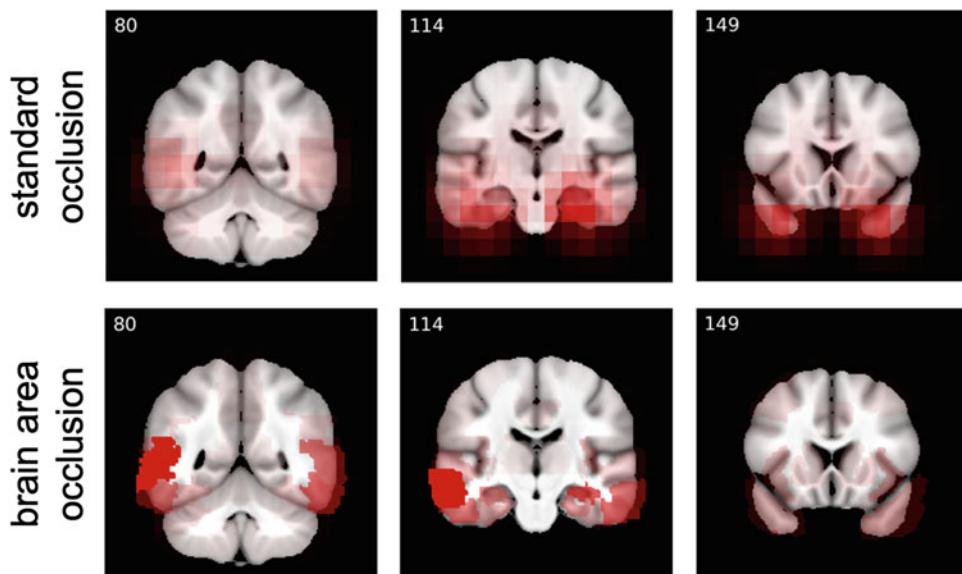


Fig. 19 Mean attribution maps obtained on demented patients. The first row corresponds to the standard and the second one to the brain area perturbation method. (Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, MLCN 2018, DLF 2018, IMIMIC 2018: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, [48], 2018)

for the first task were shared with the second one (medial temporal lobe) but that some regions were specific to the second task (parts of the parietal lobe).

Gutiérrez-Becker and Wachinger [40] adapted the standard perturbation method to a network that classified clouds of points extracted from neuroanatomical shapes of brain regions (e.g., left hippocampus) between different states of Alzheimer's disease. For the perturbation step, the authors set to 0 the coordinates of a given point x and the ones of its neighbors to then assess the relevance of the point x . This method allows easily generating and visualizing a 3D attribution map of the shapes under study.

3.4.2 Advanced Perturbation Methods

More advanced perturbation-based methods have also been used in the literature. Nigri et al. [45] compared a classical perturbation method to a swap test. The swap test replaces the classical perturbation step by a swapping step where patches are exchanged between the input brain image and a reference image chosen according to the model prediction. This exchange is possible as brain images were registered and thus brain regions are positioned in roughly the same location in each image.

Finally, Thibeau-Sutre et al. [51] used the optimized version of the perturbation method to assess the robustness of CNNs in identifying regions of interest for Alzheimer's disease detection. They applied optimized perturbations on gray matter maps extracted from T1w MR images, and the perturbation method consisted in increasing the value of the voxels to transform patients into controls. This process aimed at stimulating gray matter reconstruction to identify the most important regions that needed to be "de-atrophied" to be considered again as normal. However, they unveiled a lack of robustness of the CNN: different retrainings led to different attribution maps (shown in Fig. 20) even though the performance did not change.

3.5 Distillation Methods Applied to Neuroimaging

Distillation methods are less commonly used, but some very interesting use cases can be found in the literature on brain disorders, with methods such as LIME [44] or SHAP [33].

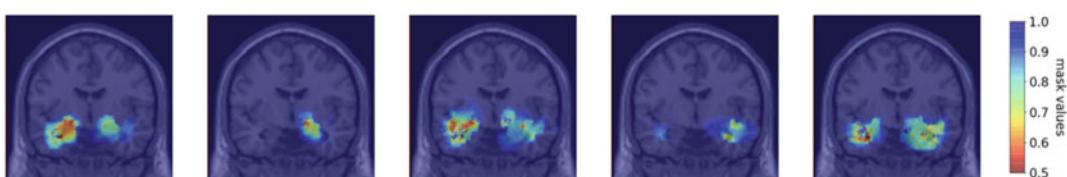


Fig. 20 Coronal view of the mean attribution masks on demented patients obtained for five reruns of the same network with the optimized perturbation method. (Adapted with permission from Medical Imaging 2020: Image Processing, [51].)

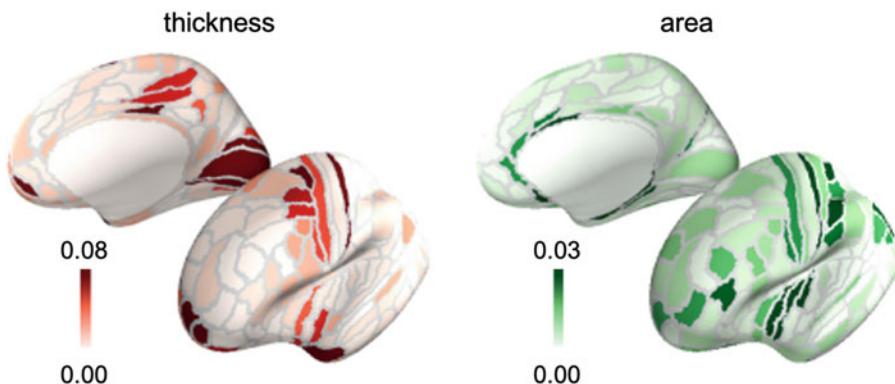


Fig. 21 Mean absolute feature importance (SHAP values) averaged across all subjects for XGBoost on regional thicknesses (red) and areas (green). (Adapted from [33] (CC BY 4.0))

Magesh et al. [44] used LIME to interpret a CNN for Parkinson's disease detection from single-photon emission computed tomography (SPECT) scans. Most of the time, the most relevant regions are the putamen and the caudate (which is clinically relevant), and some patients also showed an anomalous increase in dopamine activity in nearby areas, which is a characteristic feature of late-stage Parkinson's disease. The authors did not specify how they extracted the “super-pixels” necessary to the application of the method, though it could have been interesting to consider neuro-anatomical regions instead of creating the voxel groups with an agnostic method.

Ball et al. [33] used SHAP to obtain explanations at the individual level from three different models trained to predict participants' age from regional cortical thicknesses and areas: regularized linear model, Gaussian process regression, and XGBoost (Fig. 21). The authors exhibited a set of regions driving predictions for all models and showed that regional attention was highly correlated on average with weights of the regularized linear model. However, they showed that while being consistent across models and training folds, explanations of SHAP at the individual level were generally not correlated with feature importance obtained from the weight analysis of the regularized linear model. The authors also exemplified that the global contribution of a region to the final prediction error (“brain age delta”), even with a high SHAP value, was in general small, which indicated that this error was best explained by changes spread across several regions [33].

3.6 Intrinsic Methods Applied to Neuroimaging

3.6.1 Attention Modules

Attention modules have been increasingly used in the past couple of years, as they often allow a boost in performance while being rather easy to implement and interpret. To diagnose various brain diseases from brain CT images, Fu et al. [39] built a model integrating a “two-step attention” mechanism that selects both the most

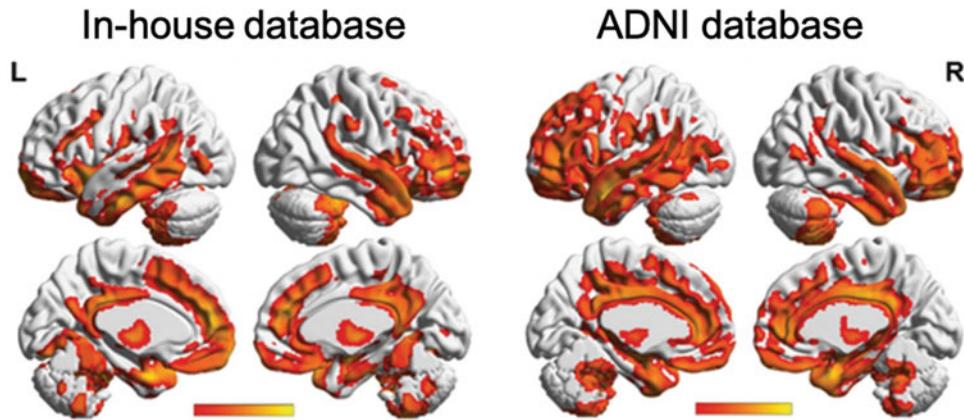


Fig. 22 Attribution maps (left, in-house database; right, ADNI database) generated by an attention mechanism module, indicating the discriminant power of various brain regions for Alzheimer's disease diagnosis. (Adapted from [42] (CC BY 4.0))

important slices and the most important pixels in each slice. The authors then leveraged these attention modules to retrieve the five most suspicious slices and highlight the areas with the more significant attention.

In their study of Alzheimer's disease, Jin et al. [42] used a 3D attention module to capture the most discriminant brain regions used for Alzheimer's disease diagnosis. As shown in Fig. 22, they obtained significant correlations between attention patterns for two independent databases. They also obtained significant correlations between regional attention scores of two different databases, which indicated a strong reproducibility of the results.

3.6.2 Modular Transparency

Modular transparency has often been used in brain imaging analysis. A possible practice consists in first generating a target probability map of a black box model, before feeding this map to a classifier to generate a final prediction, as done in [43, 46].

Qiu et al. [46] used a convolutional network to generate an attribution map from patches of the brain, highlighting brain regions associated with Alzheimer's disease diagnosis (*see* Fig. 23). Lee et al. [43] first parcellated gray matter density maps into 93 regions. For each of these regions, several deep neural networks were trained on randomly selected voxels, and their outputs were averaged to obtain a mean regional disease probability. Then, by concatenating these regional probabilities, they generated a region-wise disease probability map of the brain, which was further used to perform Alzheimer's disease detection.

The approach of Ba et al. [24] was also applied to Alzheimer's disease detection [50] (preprint). Though that work is still a pre-print, the idea is interesting as it aims at reproducing the way a radiologist looks at an MR image. The main difference with [24] is

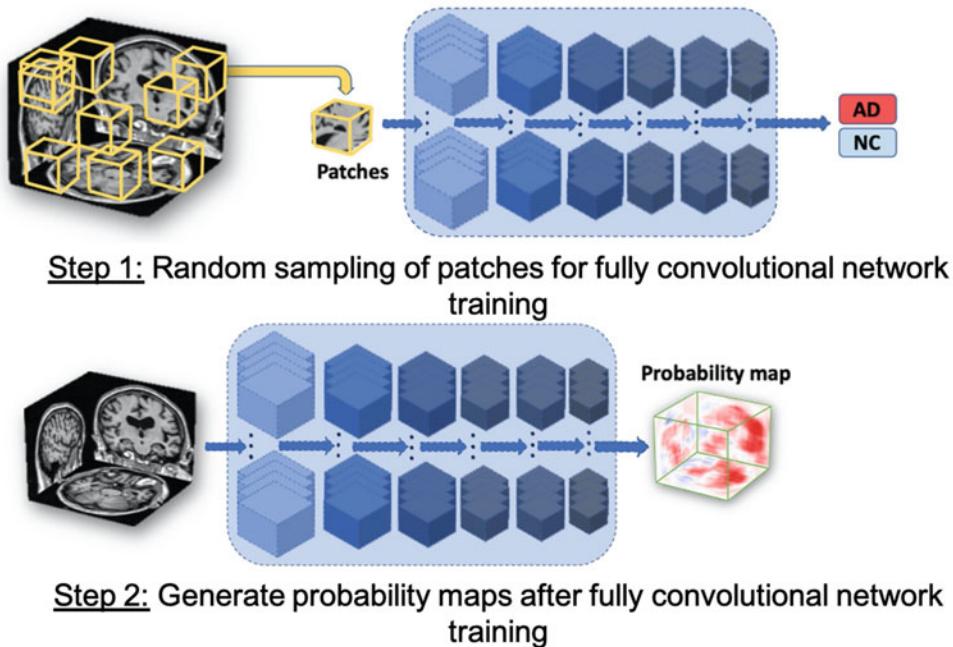


Fig. 23 Randomly selected samples of T1-weighted full MRI volumes are used as input to learn the Alzheimer's disease status at the individual level (Step 1). The application of the model to whole images leads to the generation of participant-specific disease probability maps of the brain (Step 2). (Adapted from Brain: A Journal of Neurology, 143, [46], 2020, with permission of Oxford University Press)

the initialization, as the context network does not take as input the whole image but clinical data of the participant. Then the framework browses the image in the same way as in the original paper: a patch is processed by a recurrent neural network and from its internal state the glimpse network learns which patch should be looked at next. After a fixed number of iterations, the internal state of the recurrent neural network is processed by a classification network that gives the final outcome. The whole system is interpretable as the trajectory of the locations (illustrated in Fig. 24) processed by the framework allows understanding which regions are more important for the diagnosis. However, this framework may have a high dependency to clinical data: as the initialization depends on scores used to diagnose Alzheimer's disease, the classification network may learn to classify based on the initialization only, and most of the trajectory may be negligible to assess the correct label.

Another framework, the DaniNet, proposed by Ravi et al. [47], is composed of multiple networks, each with a defined function, as illustrated in Fig. 25.

- The conditional deep autoencoder (in orange) learns to reduce the size of the slice x to a latent variable Z (encoder part) and

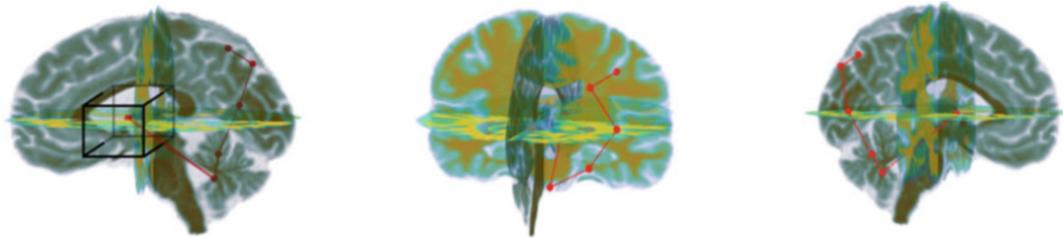


Fig. 24 Trajectory taken by the framework for a participant from the ADNI test set. A bounding box around the first location attended to is included to indicate the approximate size of the glimpse that the recurrent neural network receives; this is the same for all subsequent locations. (Adapted from [50]. Permission to reuse was kindly granted by the authors)

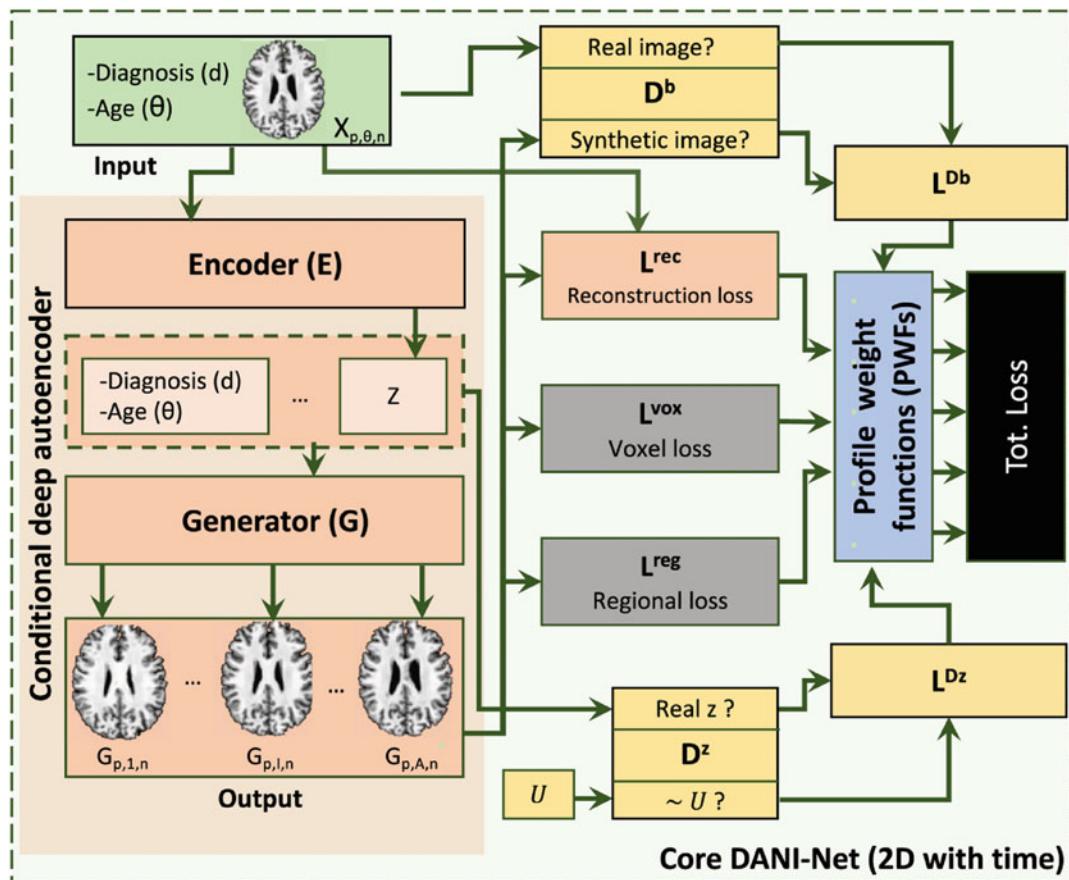


Fig. 25 Pipeline used for training the proposed DaniNet framework that aims to learn a longitudinal model of the progression of Alzheimer's disease. (Adapted from [47] (CC BY 4.0))

then to reconstruct the original image based on Z and two additional variables: the diagnosis and age (generator part). Its performance is evaluated thanks to the reconstruction loss L^{rec} .

- Discriminator networks (in yellow) either force the encoder to take temporal progression into account (D_z) or try to determine if the output of the generator are real or generated images (D_b).
- Biological constraints (in grey) force the previous generated image of the same participant to be less atrophied than the next one (voxel loss) and learn to find the diagnosis thanks to regions of the generated images (regional loss).
- Profile weight functions (in blue) aim at finding appropriate weights for each loss to compute the total loss.

The assembly of all these components allows learning a longitudinal model that characterizes the progression of the atrophy of each region of the brain. This atrophy evolution can then be visualized thanks to a neurodegeneration simulation generated by the trained model by sampling missing intermediate values.

3.7 Benchmarks Conducted in the Literature

3.7.1 Quantitative Evaluations

This section describes studies that compared several interpretability methods. We separated evaluations based on metrics from those which are purely qualitative. Indeed, even if the interpretability metrics are not mature yet, it is essential to try to measure quantitatively the difference between methods rather than to only rely on human perception, which may be biased.

Eitel and Ritter [37] tested the robustness of four methods: standard perturbation, gradientinput, guided back-propagation, and LRP. To evaluate these methods, the authors trained ten times the same model with a random initialization and generated attribution maps for each of the ten runs. For each method, they exhibited significant differences between the averaged true positives/negatives attribution maps of the ten runs. To quantify this variance, they computed the ℓ_2 -norm between the attribution maps and determined for each model the brain regions with the highest attribution. They concluded that LRP and guided back-propagation were the most consistent methods, both in terms of distance between attribution maps and most relevant brain regions. However, this study makes a strong assumption: to draw these conclusions, the network should provide stable interpretations across retrainings. Unfortunately, Thibeau-Sutre et al. [51] showed that the study of the robustness of the interpretability method and of the network should be done separately, as their network retraining was not robust. Indeed, they first showed that the interpretability method they chose (optimized perturbation) was robust according to different criteria, and then they observed that network retraining led to different attribution maps. The robustness of an interpretability method thus cannot be assessed from the protocol described in [37]. Moreover, the fact that guided back-propagation

is one of the most stable method meets the results of [6], who observed that guided back-propagation always gave the same result independently from the weights learned by a network (*see* Subheading 4.1).

Böhle et al. [34] measured the benefit of LRP with β -rule compared to guided back-propagation by comparing the intensities of the mean attribution map of demented patients and the one of cognitively normal controls. They concluded that LRP allowed a stronger distinction between these two classes than guided back-propagation, as there was a greater difference between the mean maps for LRP. Moreover, they found a stronger correlation between the intensities of the LRP attribution map in the hippocampus and the hippocampal volume than for guided back-propagation. But as [6] demonstrated that guided back-propagation has serious flaws, it does not allow drawing strong conclusions.

Nigri et al. [45] compared the standard perturbation method to a swap test (*see* Subheading 3.4) using two properties: the continuity and the sensitivity. The continuity property is verified if two similar input images have similar explanations. The sensitivity property affirms that the most salient areas in an explanation map should have the greater impact in the prediction when removed. The authors carried out experiments with several types of models, and both properties were consistently verified for the swap test, while the standard perturbation method showed a significant absence of continuity and no conclusive fidelity values [45].

Finally, Rieke et al. [48] compared four visualization methods: standard back-propagation, guided back-propagation, standard perturbation, and brain area perturbation. They computed the Euclidean distance between the mean attribution maps of the same class for two different methods and observed that both gradient methods were close, whereas brain area perturbation was different from all others. They concluded that as interpretability methods lead to different attribution maps, one should compare the results of available methods and not trust only one attribution map.

3.7.2 Qualitative Evaluations

Some works compared interpretability methods using a purely qualitative evaluation.

First, Eitel et al. [38] generated attribution maps using the LRP and gradient \odot input methods and obtained very similar results. This could be expected as it was shown that there is a strong link between LRP and gradient \odot input (*see* Subheading 2.3.2).

Dyrba et al. [36] compared DeconvNet, guided back-propagation, deep Taylor decomposition, gradient \odot input, LRP (with various rules), and Grad-CAM. The different methods roughly exhibited the same highlighted regions but with a

significant variability in focus, scatter, and smoothness, especially for the Grad-CAM method. These conclusions were derived from a visual analysis. According to the authors, LRP and deep Taylor decomposition delivered the most promising results with a highest focus and less scatter [36].

Tang et al. [49] compared two interpretability methods that seemed to have different properties: guided Grad-CAM would provide a fine-grained view of feature salience, whereas standard perturbation highlights the interplay of features among classes. A similar conclusion was drawn by Rieke et al. [48].

3.7.3 Conclusions from the Benchmarks

The most extensively compared method is LRP, and each time it has been shown to be the best method compared to others. However, its equivalence with gradient \odot input for networks using ReLU activations still questions the usefulness of the method, as gradient \odot input is much easier to implement. Moreover, the studies reaching this conclusion are not very insightful: [37] may suffer from methodological biases; [34] compared LRP only to guided back-propagation, which was shown to be irrelevant [6]; and [36] only performed a qualitative assessment.

As proposed in conclusion by Rieke et al. [48], a good way to assess the quality of interpretability methods could be to produce some form of ground truth for the attribution maps, for example, by implementing simulation models that control for the level of separability or location of differences.

4 Limitations and Recommendations

Many methods have been proposed for interpretation of deep learning models. The field is not mature yet, and none of them has become a standard. Moreover, a large panel of studies has been applied to neuroimaging data, but the value of the results obtained from the interpretability methods is often still not clear. Furthermore, many applications suffer from methodological issues, making their results (partly) irrelevant. In spite of this, we believe that using interpretability methods is highly useful, in particular to spot cases where the model exploits biases in the data set.

4.1 Limitations of the Methods

It is not often clear whether the interpretability methods really highlight features relevant to the algorithm they interpret. This way, Adebayo et al. [6] showed that the attribution maps produced by some interpretability methods (guided back-propagation and guided Grad-CAM) may not be correlated at all with the weights learned by the network during its training procedure. They prove it with a simple test called “cascading randomization.” In this test, the weights of a network trained on natural images are randomized layer per layer, until the network is fully randomized. At each step,

they produce an attribution map with a set of interpretability methods to compare it to the original ones (attribution maps produced without randomization). In the case of guided back-propagation and guided Grad-CAM, all attribution maps were identical, which means that the results of these methods were independent of the training procedure.

Unfortunately, this type of failures does not only affect interpretability methods but also the metrics designed to evaluate their reliability, which makes the problem even more complex. Tomsett et al. [52] investigated this issue by evaluating interpretability metrics with three properties:

- **Inter-rater interpretability** assesses whether a metric always rank different interpretability methods in the same way for different samples in the data set.
- **Inter-method reliability** checks that the scores given by a metric on each saliency method fluctuate in the same way between images.
- **Internal consistency** evaluates if different metrics measuring the same property (e.g., fidelity) produce correlated scores on a set of attribution maps.

They concluded that the investigated metrics were not reliable, though it is difficult to know the origin of this unreliability due to the tight coupling of model, interpretability method, and metric.

4.2 Methodological Advice

Using interpretability methods is more and more common in medical research. Even though this field is not yet mature and the methods have limitations, we believe that using an interpretability method is usually a good thing because it may spot cases where the model took decisions from irrelevant features. However, there are methodological pitfalls to avoid and good practices to adopt to make a fair and sound analysis of your results.

You should first clearly state in your paper which interpretability method you use as there exist several variants for most of the methods (*see Subheading 2*), and its parameters should be clearly specified. Implementation details may also be important: for the Grad-CAM method, attribution maps can be computed at various levels in the network; for a perturbation method, the size and the nature of the perturbation greatly influence the result. The data on which methods are applied should also be made explicit: for a classification task, results may be completely different if samples are true positives or true negatives, or if they are taken from the train or test sets.

Taking a step back from the interpretability method and especially attribution maps is fundamental as they present several limitations [34]. First, there is no ground truth for such maps, which are usually visually assessed by authors. Comparing obtained results

with the machine learning literature is a good first step, but be aware that you will most of the time find a paper to support your findings, so we suggest to look at established clinical references. Second, attribution maps are usually sensitive to the interpretability method, its parameters (e.g., β for LRP), but also to the final scale used to display maps. A slight change in one of these variables may significantly impact the interpretation. Third, an attribution map is a way to measure the impact of pixels on the prediction of a given model, but it does not provide underlying reasons (e.g., pathological shape) or explain potential interactions between pixels. A given pixel might have a low attribution when considered on its own but have a huge impact on the prediction when combined with another. Fourth, the quality of a map strongly depends on the performance of the associated model. Indeed, low-performance models are more likely to use wrong features. However, even in this case, attribution maps may be leveraged, e.g., to determine if the model effectively relies on irrelevant features (such as visual artefacts) or if there are biases in the data set [53].

One must also be very careful when trying to establish new medical findings using model interpretations, as we do not always know how the interpretability methods react when applied to correlated features. Then even if a feature seems to have no interest for a model, this does not mean that it is not useful in the study of the disease (e.g., a model may not use information from the frontal lobe when diagnosing Alzheimer's disease dementia, but this does not mean that this region is not affected by the disease).

Finally, we suggest implementing different interpretability methods to obtain complementary insights from attribution maps. For instance, using LRP in addition to the standard back-propagation method provides a different type of information, as standard back-propagation gives the sensibility of the output with respect to the input, while LRP shows the contribution of each input feature to the output. Moreover, using several metrics allows a quantitative comparison between them using interpretability metrics (*see* Subheading 2.7).

4.3 Which Method Should I Choose?

We conclude this section on how to choose an interpretability method. Some benchmarks were conducted to assess the properties of some interpretability methods compared to others (*see* Subheading 3.7). Though these are good initiatives, there are still not enough studies (and some of them suffer from methodological flaws) to draw solid conclusions. This is why we give in this section some practical advice to the reader to choose an interpretability method based on more general concepts.

Before implementing an interpretability method, we suggest reviewing the following points to help you choose carefully.

- **Implementation complexity** Some methods are more difficult to implement than others and may require substantial coding efforts. However, many of them have already been implemented in libraries or GitHub repositories (e.g., [54]), so we suggest looking online before trying to re-implement them. This is especially true for model-agnostic methods, such as LIME, SHAP, or perturbations, for which no modification of your model is required. For model-specific methods, such as back-propagation ones, the implementation will depend on the model, but if its structure is a common one (e.g., regular CNN with feature extraction followed by a classifier), it is also very likely that an adequate implementation is already available (e.g., Grad-CAM on CNN in [54]).
- **Time cost** Computation time greatly differs from one method to another, especially when input data is heavy. For instance, perturbing high dimension images is time expensive, and it would be much faster to use standard back-propagation.
- **Method parameters** The number of parameters to set varies between methods, and their choice may greatly influence the result. For instance, the patch size, the step size (distance between two patches), as well as the type of perturbation (e.g., white patches or blurry patches) must be chosen for the standard perturbation method, while the standard back-propagation does not need any parameter. Thus, without prior knowledge on the interpretability results, methods with no or only a few parameters are a good option.
- **Literature** Finally, our last piece of advice is to look into the literature to determine the methods that have commonly been used in your domain of study. A highly used method does not guarantee its quality (e.g., guided back-propagation [6]), but it is usually a good first try.

To sum up, we suggest that you choose (or at least begin with) an interpretability method that is easy to implement, time efficient, with no parameters (or only a few) to tune, and commonly used. In the context of brain image analysis, we suggest using the standard back-propagation or Grad-CAM methods. Before using a method you do not know well, you should check that other studies did not show that this method is not relevant (which is the case for guided back-propagation or guided Grad-CAM) or that it is not equivalent to another method (e.g., LRP on networks with ReLU activation layers and gradient \odot input).

Regarding interpretability metrics, there is no consensus in the community as the field is not mature yet. General advice would be to use different metrics and confront them to human observers, taking, for example, the methodology described in [1].

5 Conclusion

Interpretability of machine learning models is an important topic, in particular in the medical field. First, this is a natural need expressed by clinicians who are potential users of medical decision support systems. Moreover, it has been shown in many occasions that models with high performance can actually be using irrelevant features. This is dangerous because it means that they are exploiting biases in the training data sets and thus may dramatically fail when applied to new data sets or deployed in clinical routine.

Interpretability is a very active field of research and many approaches have been proposed. They have been extensively applied in neuroimaging and very often allowed highlighting clinically relevant regions of the brain that were used by the model. However, comparative benchmarks are not entirely conclusive, and it is currently not clear which approach is the most adapted for a given aim. In other words, it is very important to keep in mind that the field of interpretability is not yet mature. It is not yet clear which are the best methods or even if the most widely used approaches will still be considered a standard in the near future.

That being said, we still strongly recommend that a classification or regression model be studied with at least one interpretability method. Indeed, evaluating the performance of the model is not sufficient in itself, and the additional use of an interpretation method may allow detecting biases and models that perform well but for bad reasons and thus would not generalize to other settings.

Acknowledgements

The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6).

Appendices

A Short Reminder on Network Training Procedure

During the training phase, a neural network updates its weights to make a series of inputs match with their corresponding target labels:

1. *Forward pass* The network processes the input image to compute the output value.

2. *Loss computation* The difference between the true labels and the output values is computed according to a criterion (cross-entropy, mean squared error...). This difference is called the loss and should be as low as possible.
3. *Backward pass* For each learnable parameter of the network, the gradients with respect to the loss are computed.
4. *Weight update* Weights are updated according to the gradients and an optimizer rule (stochastic gradient descent, Adam, Adadelta...).

As a network is a composition of functions, the gradients of the weights of a layer l with respect to the loss can be easily obtained according to the values of the gradients in the following layers. This way of computing gradients layer per layer is called back-propagation.

B Description of the Main Brain Disorders Mentioned in the Reviewed Studies

This appendix aims at shortly presenting the diseases considered by the studies reviewed in Subheading 3.

The majority of the studies focused on the classification of Alzheimer's disease (AD), a neurodegenerative disease of the elderly. Its pathological hallmarks are senile plaques formed by amyloid- β protein and neurofibrillary tangles that are tau protein aggregates. Both can be measured in vivo using either PET imaging or CSF biomarkers. Several other biomarkers of the disease exist. In particular, atrophy of gray and white matter measured from T1w MRI is often used, even though it is not specific to AD. There is strong and early atrophy in the hippocampi that can be linked to the memory loss, even though other clinical signs are found and other brain areas are altered. The following diagnosis statuses are often used:

- **AD** refers to demented patients.
- **CN** refers to cognitively normal participants.
- **MCI** refers to patients with mild cognitive impairment (they have an objective cognitive decline, but it is not sufficient yet to cause a loss of autonomy).
- **Stable MCI** refers to MCI patients who stayed stable during a defined period (often three years).
- **Progressive MCI** refers to MCI patients who progressed to Alzheimer's disease during a defined period (often three years).

Most of the studies analyzed T1w MRI data, except [49] where the patterns of amyloid- β in the brain are studied.

Frontotemporal dementia is another neurodegenerative disease in which the neuronal loss dominates in the frontal and temporal lobes. Behavior and language are the most affected cognitive functions.

Parkinson's disease is also a neurodegenerative disease. It primarily affects dopaminergic neurons in the substantia nigra. A commonly used neuroimaging technique to detect this loss of dopaminergic neurons is the SPECT, as it uses a ligand that binds to dopamine transporters. Patients are affected by different symptoms linked to motor faculties such as tremor, slowed movements, and gait disorder but also sleep disorder, depression, and other symptoms.

Multiple sclerosis is a demyelinating disease with a neurodegenerative component affecting younger people (it begins between the ages of 20 and 50). It causes demyelination of the white matter in the brain (brain stem, basal ganglia, tracts near the ventricles), optic nerve, and spinal cord. This demyelination results in autonomic, visual, motor, and sensory problems.

Intracranial hemorrhage may result from a physical trauma or non-traumatic causes such as a ruptured aneurysm. Different subtypes exist depending on the location of the hemorrhage.

Autism is a spectrum of neurodevelopmental disorders affecting social interaction and communication. Diagnosis is done based on clinical signs (behavior), and the patterns that may exist in the brain are not yet reliably described as they overlap with the neurotypical population.

Some brain characteristics that may be related to brain disorders and detected in CT scans were considered in the data set CQ500:

- **Midline Shift** is a shift of the center of the brain past the center of the skull.
- **Mass Effect** is caused by the presence of an intracranial lesion (e.g., a tumor) that is compressing nearby tissues.
- **Calvarial Fractures** are fractures of the skull.

Finally, one study [33] learned to predict the age of cognitively normal patients. Such algorithm can help in diagnosing brain disorders as patients will have a greater brain age than their chronological age, and then it establishes that a participant is not in the normal distribution.

References

1. Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you?: Explaining the predictions of any Classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining – KDD '16, ACM Press, San Francisco, pp 1135–1144. <https://doi.org/10.1145/2939672.2939778>
2. Fong RC, Vedaldi A (2017) Interpretable explanations of black boxes by meaningful perturbation. In: 2017 IEEE international conference on computer vision (ICCV), pp 3449–3457. <https://doi.org/10.1109/ICCV.2017.371>
3. DeGrave AJ, Janizek JD, Lee SI (2021) AI for radiographic COVID-19 detection selects

- shortcuts over signal. *Nat Mach Intell* 3(7): 610–619. <https://doi.org/10.1038/s42256-021-00338-7>
4. Lipton ZC (2018) The mythos of model interpretability. *Commun ACM* 61(10):36–43. <https://doi.org/10.1145/3233231>
 5. Xie N, Ras G, van Gerven M, Doran D (2020) Explainable deep learning: a field guide for the uninitiated. *arXiv:200414545 [cs, stat]* 2004.14545
 6. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. In: *Advances in Neural Information Processing Systems*, pp 9505–9515
 7. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) *Advances in Neural information processing systems*, vol 25. Curran Associates, pp 1097–1105
 8. Voss C, Cammarata N, Goh G, Petrov M, Schubert L, Egan B, Lim SK, Olah C (2021) Visualizing weights. *Distill* 6(2):e00024.007. <https://doi.org/10.23915/distill.00024.007>
 9. Olah C, Mordvintsev A, Schubert L (2017) Feature visualization. *Distill* 2(11):e7. <https://doi.org/10.23915/distill.00007>
 10. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv:13126034 [cs]* 1312.6034
 11. Shrikumar A, Greenside P, Shcherbina A, Kundaje A (2017) Not just a black box: learning important features through propagating activation differences. *arXiv:160501713 [cs]* 1605.01713
 12. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M (2014) Striving for Simplicity: the all convolutional net. *arXiv:14126806 [cs]* 1412.6806
 13. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2015) Learning deep features for discriminative localization. *arXiv:151204150 [cs]* 1512.04150
 14. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *2017 IEEE international conference on computer vision (ICCV)*, pp 618–626. <https://doi.org/10.1109/ICCV.2017.74>
 15. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS One* 10(7):e0130140. <https://doi.org/10.1371/journal.pone.0130140>
 16. Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR (2017) Evaluating the visualization of what a deep neural network has learned. *IEEE Trans Neural Netw Learn Syst* 28(11):2660–2673. <https://doi.org/10.1109/TNNLS.2016.2599820>
 17. Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR (2017) Explaining non-linear classification decisions with deep Taylor decomposition. *Pattern Recogn* 65:211–222. <https://doi.org/10.1016/j.patcog.2016.11.008>
 18. Montavon G, Samek W, Müller KR (2018) Methods for interpreting and understanding deep neural networks. *Digit Signal Process* 73:1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>
 19. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) *Computer vision – ECCV 2014. Lecture notes in computer science*. Springer, Berlin, pp 818–833
 20. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: *Proceedings of the 31st international conference on neural information processing systems, NIPS’17*. Curran Associates, Red Hook, pp 4768–4777
 21. Frosst N, Hinton G (2017) Distilling a Neural network into a soft decision tree. *arXiv:171109784 [cs, stat]* 1711.09784
 22. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y (2016) Show, attend and tell: neural image caption generation with visual attention. *arXiv:150203044 [cs]* 1502.03044
 23. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual attention network for image classification. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, Honolulu, pp 6450–6458. <https://doi.org/10.1109/CVPR.2017.683>
 24. Ba J, Mnih V, Kavukcuoglu K (2015) Multiple object recognition with visual attention. *arXiv:14127755 [cs]* 1412.7755
 25. Yeh CK, Hsieh CY, Suggala A, Inouye DI, Ravikumar PK (2019) On the (In)fidelity and sensitivity of explanations. In: Wallach H, Larochelle H, Beygelzimer A, d\text{\\}textquotesingle Alché-Buc F, Fox E, Garnett R (eds) *Advances in neural information processing systems*, vol 32. Curran Associates, pp 10967–10978

26. Cecotti H, Gräser A (2011) Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Trans on Pattern Anal Mach Intell* 33(3): 433–445. <https://doi.org/10.1109/TPAMI.2010.125>
27. Oh K, Chung YC, Kim KW, Kim WS, Oh IS (2019) Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning. *Sci Rep* 9(1):1–16. <https://doi.org/10.1038/s41598-019-54548-6>
28. Abrol A, Bhattacharai M, Fedorov A, Du Y, Plis S, Calhoun V (2020) Deep residual learning for neuroimaging: an application to predict progression to Alzheimer's disease. *J Neurosci Methods* 339:108701. <https://doi.org/10.1016/j.jneumeth.2020.108701>
29. Biffi C, Cerrolaza J, Tarroni G, Bai W, De Marvao A, Oktay O, Ledig C, Le Folgoc L, Kamnitsas K, Doumou G, Duan J, Prasad S, Cook S, O'Regan D, Rueckert D (2020) Explainable anatomical shape analysis through deep Hierarchical generative models. *IEEE Trans Med Imaging* 39(6):2088–2099. <https://doi.org/10.1109/TMI.2020.2964499>
30. Martinez-Murcia FJ, Ortiz A, Gorri JM, Ramirez J, Castillo-Barnes D (2020) Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders. *IEEE J Biomed Health Inf* 24(1):17–26. <https://doi.org/10.1109/JBHI.2019.2914970>
31. Leming M, Gorri JM, Suckling J (2020) Ensemble deep learning on large, mixed-site fMRI datasets in autism and other tasks. *Int J Neural Syst* 2050012. <https://doi.org/10.1142/S0129065720500124>, 2002.07874
32. Bae J, Stocks J, Heywood A, Jung Y, Jenkins L, Katsaggelos A, Popuri K, Beg MF, Wang L (2019) Transfer learning for predicting conversion from mild cognitive impairment to dementia of Alzheimer's type based on 3D-convolutional neural network. *bioRxiv*. <https://doi.org/10.1101/2019.12.20.884932>
33. Ball G, Kelly CE, Beare R, Seal ML (2021) Individual variation underlying brain age estimates in typical development. *Neuroimage* 235:118036. <https://doi.org/10.1016/j.neuroimage.2021.118036>
34. Böhle M, Eitel F, Weygandt M, Ritter K, on botADNI (2019) Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Front Aging Neurosci* 10(JUL). <https://doi.org/10.3389/fnagi.2019.00194>
35. Burduja M, Ionescu RT, Verga N (2020) Accurate and efficient intracranial hemorrhage detection and subtype classification in 3D CT scans with convolutional and long short-term memory neural networks. *Sensors* 20(19): 5611. <https://doi.org/10.3390/s20195611>
36. Dyrba M, Pallath AH, Marzban EN (2020) Comparison of CNN visualization methods to aid model interpretability for detecting Alzheimer's disease. In: Tolxdorff T, Deserno TM, Handels H, Maier A, Maier-Hein KH, Palm C (eds) *Bildverarbeitung für die Medizin 2020*, Springer Fachmedien, Wiesbaden, Informatik aktuell, pp 307–312. https://doi.org/10.1007/978-3-658-29267-6_68
37. Eitel F, Ritter K (2019) Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. In: *Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Lecture notes in computer science*. Springer, Cham, pp 3–11. https://doi.org/10.1007/978-3-030-33850-3_1
38. Eitel F, Soehler E, Bellmann-Strobl J, Brandt AU, Ruprecht K, Giess RM, Kuchling J, Asseyer S, Weygandt M, Haynes JD, Scheel M, Paul F, Ritter K (2019) Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *Neuroimage: Clinical* 24:102003. <https://doi.org/10.1016/j.nicl.2019.102003>
39. Fu G, Li J, Wang R, Ma Y, Chen Y (2021) Attention-based full slice brain CT image diagnosis with explanations. *Neurocomputing* 452: 263–274. <https://doi.org/10.1016/j.neucom.2021.04.044>
40. Gutiérrez-Becker B, Wachinger C (2018) Deep multi-structural shape analysis: application to neuroanatomy. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, LNCS, vol 11072, pp 523–531. https://doi.org/10.1007/978-3-030-00931-1_60
41. Hu J, Qing Z, Liu R, Zhang X, Lv P, Wang M, Wang Y, He K, Gao Y, Zhang B (2021) Deep learning-based classification and voxel-based visualization of frontotemporal dementia and Alzheimer's disease. *Front Neurosci* 14. <https://doi.org/10.3389/fnins.2020.626154>
42. Jin D, Zhou B, Han Y, Ren J, Han T, Liu B, Lu J, Song C, Wang P, Wang D, Xu J, Yang Z, Yao H, Yu C, Zhao K, Wintermark M, Zuo N,

- Zhang X, Zhou Y, Zhang X, Jiang T, Wang Q, Liu Y (2020) Generalizable, reproducible, and neuroscientifically interpretable imaging biomarkers for Alzheimer's disease. *Adv Sci* 7(14):2000675. <https://doi.org/10.1002/advs.202000675>
43. Lee E, Choi JS, Kim M, Suk HI (2019) Alzheimer's disease neuroimaging initiative toward an interpretable Alzheimer's disease diagnostic model with regional abnormality representation via deep learning. *NeuroImage* 202: 116113. <https://doi.org/10.1016/j.neuroimage.2019.116113>
44. Magesh PR, Myloth RD, Tom RJ (2020) An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. *Comput Biol Med* 126:104041. <https://doi.org/10.1016/j.combiomed.2020.104041>
45. Nigri E, Ziviani N, Cappabianco F, Antunes A, Veloso A (2020) Explainable deep CNNs for MRI-based diagnosis of Alzheimer's disease. In: Proceedings of the international joint conference on neural networks. <https://doi.org/10.1109/IJCNN48605.2020.9206837>
46. Qiu S, Joshi PS, Miller MI, Xue C, Zhou X, Karjadi C, Chang GH, Joshi AS, Dwyer B, Zhu S, Kaku M, Zhou Y, Alderazi YJ, Swaminathan A, Kedar S, Saint-Hilaire MH, Auerbach SH, Yuan J, Sartor EA, Au R, Kolachalam VB (2020) Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain: J Neurol* 143(6):1920–1933. <https://doi.org/10.1093/brain/awaa137>
47. Ravi D, Blumberg SB, Ingala S, Barkhof F, Alexander DC, Oxtoby NP (2022) Degenerative adversarial neuroimage nets for brain scan simulations: application in ageing and dementia. *Med Image Anal* 75:102257. <https://doi.org/10.1016/j.media.2021.102257>
48. Rieke J, Eitel F, Weygandt M, Haynes JD, Ritter K (2018) Visualizing convolutional networks for MRI-based diagnosis of Alzheimer's disease. In: Understanding and interpreting machine learning in medical image computing applications. Lecture notes in computer science. Springer, Cham, pp 24–31. https://doi.org/10.1007/978-3-030-02628-3_3
49. Tang Z, Chuang KV, DeCarli C, Jin LW, Beckett L, Keiser MJ, Dugger BN (2019) Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. *Nat Commun* 10(1):1–14. <https://doi.org/10.1038/s41467-019-10212-1>
50. Wood D, Cole J, Booth T (2019) NEURO-DRAM: a 3D recurrent visual attention model for interpretable neuroimaging classification. arXiv:191004721 [cs, stat] 1910.04721
51. Thibeau-Sutre E, Colliot O, Dormont D, Burgos N (2020) Visualization approach to assess the robustness of neural networks for medical image classification. In: Medical imaging 2020: image processing, international society for optics and photonics, vol 11313, p 11313J. <https://doi.org/10.1117/12.2548952>
52. Tomsett R, Harborne D, Chakraborty S, Gurram P, Preece A (2020) Sanity checks for saliency metrics. *Proc AAAI Conf Artif Intell* 34(04):6021–6029. <https://doi.org/10.1609/aaai.v34i04.6064>
53. Lapuschkin S, Binder A, Montavon G, Muller KR, Samek W (2016) Analyzing classifiers: fisher vectors and deep neural networks. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Las Vegas, pp 2912–2920. <https://doi.org/10.1109/CVPR.2016.318>
54. Ozbulak U (2019) Pytorch cnn visualizations. <https://github.com/utkuozbulak/pytorch-cnn-visualizations>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third-party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 23

A Regulatory Science Perspective on Performance Assessment of Machine Learning Algorithms in Imaging

Weijie Chen, Daniel Krainak, Berkman Sahiner, and Nicholas Petrick

Abstract

This chapter presents a regulatory science perspective on the assessment of machine learning algorithms in diagnostic imaging applications. Most of the topics are generally applicable to many medical imaging applications, while brain disease-specific examples are provided when possible. The chapter begins with an overview of US FDA's regulatory framework followed by assessment methodologies related to ML devices in medical imaging. Rationale, methods, and issues are discussed for the study design and data collection, the algorithm documentation, and the reference standard. Finally, study design and statistical analysis methods are overviewed for the assessment of standalone performance of ML algorithms as well as their impact on clinicians (i.e., reader studies). We believe that assessment methodologies and regulatory science play a critical role in fully realizing the great potential of ML in medical imaging, in facilitating ML device innovation, and in accelerating the translation of these technologies from bench to bedside to the benefit of patients.

Key words Machine learning, Performance assessment, Standalone performance, Reader study, Statistical analysis plan, Regulatory science

1 Introduction

Machine learning (ML) technologies are being developed at an ever-increasing pace in a variety of medical imaging applications [1]. Particularly in brain imaging, the past decade has witnessed a spectacular growth of ML development for the diagnosis, prognosis, and treatment of brain disorders [2]. One of the ultimate goals of these developments is to translate safe and effective technologies to the clinic to benefit patients. Regulatory oversight plays a key role in this translation. The mission of the Center for Devices and Radiological Health (CDRH) at the US Food and Drug Administration (US FDA) is to “assure that patients and providers have timely and continued access to safe, effective, and high-quality

medical devices.”¹ This chapter discusses performance assessment of machine learning algorithms in imaging applications from a regulatory science perspective. Regulatory science is the science of developing new tools, standards, and approaches to assess the safety, efficacy, quality, and performance of all FDA-regulated products.²

We begin with clarifications of the scope of this chapter. First, following an overview of the US FDA’s regulatory framework for medical imaging and related ML devices, the primary topics we discuss are about concepts, basic principles, and methods for performance assessment of ML algorithms in the arena of *regulatory science* but not *regulatory policy*. As such, these topics are not necessarily relevant to every regulatory submission. The question of which components should be included in a specific regulatory submission is a regulatory decision depending on factors such as the risk of the device, impact on clinical practice, complexity of the technology, precedents, and so on and is beyond the scope of this chapter. Second, the topics are *selected* based on our experience and expertise but are not intended to be comprehensive. For example, software engineering and cybersecurity are important aspects of ML devices but are beyond the scope of this chapter. Third, as discussed in earlier chapters of this book, ML algorithms are developed for both *imaging* and *non-imaging* modalities for treating brain disorders. We focus on imaging applications. Moreover, while this book is on brain disorders, most of the discussions in this chapter are applicable to ML algorithms in general imaging applications unless noted otherwise. Lastly, while the assessment methods are well established to the best of our knowledge at the time of writing, we acknowledge that ML techniques and assessment methodologies are active areas of research and better methods may become available and adopted by researchers, developers, and regulatory agencies alike in the future. To give the readers a more specific sense of the scope of applications that are relevant to our discussions, we reviewed, via the American College of Radiology (ACR) and FDA public databases, some ML devices for brain disorders that were authorized by the FDA in recent years and summarized major scope characteristics including the imaging modalities, functionalities, and types of ML algorithms (*see Table 1*).

The rest of the chapter begins with an overview of US FDA’s regulatory framework followed by topics on assessment methodologies related to ML devices in medical imaging. Rationale, methods, and issues are discussed for study design and data collection

¹ <https://www.fda.gov/about-fda/center-devices-and-radiological-health/cdrh-mission-vision-and-shared-values>

² <https://www.fda.gov/science-research/science-and-research-special-topics/advancing-regulatory-science>

Table 1
Summary characteristics of exemplar FDA-cleared ML devices for brain disorders

Modality	CT (contrast or non-contrast), CTA, MRI, PET, SPECT
Functionality	Triage and notification (e.g., for intracranial hemorrhage); segmentation, quantification, and feature measurements; analysis and visualization; computer-aided diagnosis; denoising, enhancement; auto-contouring/segmentation of organs at risk or tumors for radiation therapy of head and neck tumors
ML algorithms	Hand-crafted feature extraction and computerized classifiers; deep learning neural networks

CT computed tomography, *CTA* computed tomography angiography, *MRI* magnetic resonance imaging, *PET* positron emission tomography, *SPECT* single photon emission computed tomography. Summary based on a sampled review of public databases at ACR (<https://models.acrdsi.org/>) and FDA (<https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices>) websites. The table aims to give a general overview of the scope of devices available. For specific devices that work for a certain imaging modality with certain functionalities, please refer to the cited databases

(Subheading 3), algorithm documentation (Subheading 4), and reference standard (Subheading 5). Finally, performance assessment methodologies are overviewed including the standalone performance assessment of ML algorithms (Subheading 6), assessment of ML algorithms in the hands of clinicians (i.e., reader studies; Subheading 7), and general considerations for the statistical analysis (Subheading 8). The relationships among these topics are illustrated in Fig. 1. Performance assessment of ML devices is necessary in both premarket and postmarket environments. Premarket studies are for the assessment of safety and effectiveness before the device is authorized for marketing by a regulatory body. Some premarket studies are used in the context of device development to refine and iterate on device design. Other premarket studies are intended for review by regulatory bodies to help assess the safety and effectiveness prior to marketing authorization. Post-market studies are for clinical use and epidemiology, maintenance, and modifications. The selected topics to be discussed in this chapter belong to premarket performance assessment.

2 Regulatory Framework

CDRH Learn³ provides readers an excellent resource to better understand overall medical device regulation.

2.1 Overview

The US FDA classifies medical devices into three classes, Classes I, II, and III. The classification determines the extent of regulatory controls necessary to provide reasonable assurance of the safety and

³ <https://www.fda.gov/training-and-continuing-education/cdrh-learn>

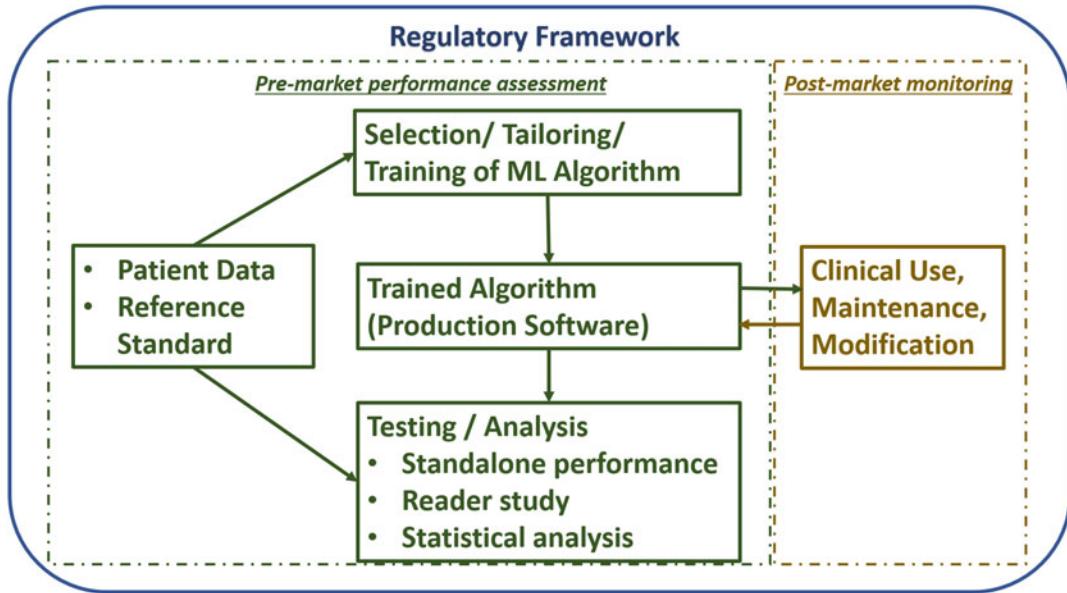


Fig. 1 ML performance assessment methods in the context of US FDA's regulatory framework

effectiveness of the device. The device classification tends to increase with increasing degree of risk, and the appropriate types of controls applicable to the device depend on the device classification. There are three types of regulatory controls: general controls, special controls, and premarket authorization requirements. General controls include the basic provisions applicable to medical devices of the Food, Drug, and Cosmetic Act and apply to all medical devices. They include provisions that relate to adulteration; misbranding; device registration and listing; premarket notification; banned devices; notification, including repair, replacement, or refund; records and reports; restricted devices; and good manufacturing practices.⁴ Special controls apply to Class II devices and are published in the Code of Federal Regulations under the specific device type. Some examples of special controls include labeling, testing, design specifications, software life cycle documentation activities, and usability assessments.

The US FDA requirements for premarket submissions differ between the device classes. To receive FDA approval, sponsors of Class III devices, generally considered the highest risk devices, must demonstrate a reasonable assurance of safety and effectiveness. Sponsors of Class I and II device must demonstrate substantial equivalence between their new device and a legally marketed device through the premarket notification process (i.e., the 510 [k] Program), unless the product class is exempt from premarket

⁴ <https://www.fda.gov/medical-devices/regulatory-controls/general-controls-medical-devices>

notification. Substantial equivalence is a comparative analysis that includes a comparison of the intended use, technological characteristics, and performance testing. For device classifications that include defined special controls (generally published in the Code of Federal Regulations or in an order granting a request for reclassification), the sponsor must also demonstrate that they have fulfilled all the necessary special controls as part of the premarket notification process and to avoid marketing an adulterated or misbranded device.

The De Novo classification process is a pathway to Class I or Class II classification for medical devices for which general controls or general and special controls provide a reasonable assurance of safety and effectiveness, but for which there is no legally marketed predicate device [3]. Devices of a new type that FDA has not previously classified are “automatically” or “statutorily” classified into Class III by the FD&C Act, regardless of the level of risk they pose or the ability of general and special controls to assure safety and effectiveness. Section 513(f)(2) of the FD&C Act allows manufacturers to submit a De Novo request to FDA for devices “automatically” classified into Class III by operation of Section 513(f)(1). In essence, a De Novo is a request for classification for a novel device that would otherwise be classified as a Class III device. During review of a De Novo request, the FDA evaluates whether general controls or general and special controls are adequate to provide a reasonable assurance of safety and effectiveness for the identified classification of the device.

FDA regulates products based upon the device characteristics (e.g., what is it? what does it do?) and the intended use of the device. The submission type and performance data necessary to obtain marketing authorization depends on the device classification, technological characteristics, and intended use. Understanding the technological characteristics is often a more straightforward exercise compared to the determination of the intended use of the product when attempting to determine the appropriate regulatory pathway and necessary supporting data. Intended use means the general purpose of the device or its function and encompasses the indications for use [4]. The indications for use, as defined in 21 CFR 814.20(b)(3)(i), describes the disease or condition the device will diagnose, treat, prevent, cure, or mitigate, including a description of the patient population for which the device is intended. The intended use of a device is one criterion that determines whether a device can be cleared for marketing through the 510(k) process or must be evaluated as a Class III device (premarket approval) or, if appropriate, a De Novo request. Section 513(i)(1)(E)(i) of the FD&C Act provides that the FDA’s determination of intended use of a device “shall be based upon the proposed labeling.” A device may have a variety of different indications for use and

intended uses (e.g., output a measurement for users, identify patients eligible for a particular treatment, estimate prognostic cancer risk, or predict a patient's response to therapy). The data needed to support these different intended uses and indications are different.

2.2 Imaging Device Regulation

A majority of medical image processing devices have been classified as Class II devices. Most of the software-only devices or software as a medical device that are intended for image processing have been classified under 21 CFR 892.2050 as picture archiving and communications systems. On April 19, 2021, FDA updated the name of the regulation 21 CFR 892.2050 to "medical image management and processing system." There are no published, mandatory specific special controls related to software-only devices classified under 21 CFR 892.2050, and therefore, the primary resource to understand the legal requirements for performance data associated with these devices is the comparative standard of substantial equivalence as described in detail in the guidance document on the 510 (k) Program [4]. In contrast, several devices more recently classified under the De Novo pathway have specific special controls that manufacturers marketing such devices must adhere to.

Devices originally classified via the De Novo pathway often include special controls defined in the CFR describing requirements for manufacturers of these devices. Devices that may implement machine learning that include software or software-only devices must adhere to the special controls defined in the specific regulations associated with the appropriate device class. The classification with the associated special controls is published with a Federal Register notice and appears in the Electronic Code of Federal Regulations (eCFR).⁵ A De Novo classification, including any special control, is effective on the date the order letter is issued granting the De Novo request [3]. For the specific examples cited below, the De Novo submission (DEN number) is cited for classifications that have not been published in CFR at the time of writing, and the associated order with special controls may be found by searching FDA's De Novo database.⁶ Examples include:

- 21 CFR 870.2785 (DEN200019): Software for optical camera-based measurement of pulse rate, heart rate, breathing rate, and/or respiratory rate
- 21 CFR 870.2790 (DEN200038): Hardware and software for optical camera-based measurement of pulse rate, heart rate, breathing rate, and/or respiratory rate

⁵ <https://www.ecfr.gov/cgi-bin/ECFR?page=browse>

⁶ <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm>

- 21 CFR 876.1520 (DEN200055): Gastrointestinal lesion software detection system
- 21 CFR 892.2060 (DEN170022): Radiological computer-assisted diagnostic software for lesions suspicious of cancer
- 21 CFR 892.2070: Medical image analyzer
- 21 CFR 892.2080 (DEN170073): Radiological computer-aided triage and notification software
- 21 CFR 892.2090 (DEN180005): Radiological computer-assisted detection and diagnosis software
- 21 CFR 892.2100 (DEN190040): Radiological acquisition and/or optimization guidance system

The special controls associated with these regulations are intended to mitigate the risks to health associated with these types of devices. As such, many of the special controls included in these classifications relate directly to elements associated with machine learning-based software devices intended for use in diagnostics. For example, several of the regulations include special controls related to the description of the image analysis algorithm (e.g., 21 CFR 892.2060(b)(1)(i), 21 CFR 870.2785(1), 21 CFR 876.1520(5)). Many others specify elements of the performance testing and characterization. Often included in these regulations (e.g., 21 CFR 892.2060, 21 CFR 892.2070) are special controls that indicate performance must demonstrate that the device provides improved performance on a particular diagnostic task (e.g., detection, diagnosis). For new devices, these requirements generally mean FDA will require both standalone testing characterizing device performance and clinical testing demonstrating diagnostic improvement in the intended use population. For devices implementing machine learning algorithms to estimate other physiologic characteristics, standalone and clinical testing may also be required (e.g., 21 CFR 870.2785). In addition, these regulations may include special controls related to describing the expected performance of the device. Requirements associated with communicating expected device performance in labeling help to (a) mitigate the risks associated with the device and (b) communicate expectations for performance for similar devices to future device developers.

CDRH is statutorily mandated to consider the least burdensome approach to regulatory requirements or decisions. Alternative methods, data sources, real-world evidence, nonclinical data, and other means to meet regulatory requirements may be considered and accepted, when appropriate. FDA encourages innovative approaches to device design as well as mechanisms to address regulatory requirements, when appropriate. FDA takes a benefit-

risk approach to novel devices [5] and to devices with different technological characteristics [6].

CDRH provides opportunities for developers to request feedback and meet with FDA staff to obtain FDA feedback prior to an intended premarket submission [7]. These interactions tend to focus on a particular device and questions relevant to a planned future regulatory submission and may include questions about testing protocols, proposed labeling, regulatory pathways, and design and performance of clinical studies and acceptance criteria.

Device developers need to be aware of all regulatory requirements throughout a product's life cycle including investigational device requirements (e.g., 21 CFR 812), premarket requirements, postmarket requirements (e.g., 21 CFR 820), and surveillance requirements. While this chapter focuses on the premarket and performance assessment of devices, we remind the reader that regulatory requirements throughout the device life cycle should be considered.

3 Study Design and Data Collection

This section aims at summarizing general considerations for study design and data collection for the assessment of ML algorithms in imaging. The specific topics we focus on in this section include study objectives, pilot and pivotal studies, and issues related to data collection, including dataset mismatch and bias. Other study design considerations, such as selection of a reference standard, selection of a performance metric, and data analysis plans, are discussed in later sections.

3.1 Study Objectives

The first consideration in study design is the objective of the study. A general principle is that the study design should aim at generating data to support what the ML algorithm claims to accomplish. The required data are closely related to the intended use of the device, including the target patient population. Important considerations include the significance of information provided by an ML algorithm to a healthcare decision, the state of the healthcare situation or condition that the algorithm addresses, and how the ML algorithm is intended to be integrated into the current standard of care. Examples of study objectives for ML algorithms include standalone performance characterization, standalone performance comparison with another algorithm or device, performance characterization of human users when equipped with the algorithm, and performance comparison of human users with and without the algorithm.

3.2 Pilot and Pivotal Studies

For the purpose of this chapter, a pivotal study is defined as a definitive study in which evidence is gathered to support the safety and effectiveness evaluation of a medical device for its intended use. A pivotal study is the key formal performance assessment of ML devices in medical imaging, and the design of a pivotal study is often the culmination of a significant amount of previous work. An often overlooked, important step toward the design of a pivotal study is a pilot (or exploratory) study. Pilot studies may include different phases, including those that demonstrate the engineering proof of concept, those that lead to a better understanding of the mechanisms involved, those that may lead to iterative improvements in performance, and those that yield essential information for designing a pivotal study. When a pilot study involves patients, sample size is typically small, and data are often conveniently acquired rather than representative of an intended population [8]. Such pilot studies provide information about the estimates of the effect size and variance components that are critical for estimating the sample size for a pivotal study. In addition, a pilot study can uncover basic issues in data collection, including issues about missing or incomplete data and poor imaging protocols. For pivotal studies that include clinicians (typically radiologists or pathologists who interpret images when equipped with the ML algorithm), a pilot study can reveal poor reading protocols and poor reader training [8]. Running one or more pilot studies is therefore highly advisable prior to the design of a pivotal study.

3.3 Data Collection

An important prerequisite for a study that supports the claims of an ML algorithm is that the data collection process should allow the replication of the conclusions drawn from this particular study by independent studies in the future. In this regard, the composition and independence of training and test datasets and dataset representativeness are central issues.

3.3.1 Training and Test Datasets

Training data are defined as the set of patient-related attributes (raw data, images, and other associated information) used for inferring a function between these attributes and the desired output for the ML algorithm. During training, investigators may explore different algorithm architectures for this function and fine-tune the parameters of a selected architecture. The algorithm designer can also partition this data into different sets for preliminary (or exploratory) performance analysis, utilizing, for example, cross-validation techniques [9]. Typically, these cross-validation results are used for further model development, model selection, and hyperparameter tuning. In other words, cross-validation is typically used as an informative step before the ML algorithm is finalized. In many machine learning texts, a subset of data left out

for certain parts of algorithm design (e.g., tuning hyperparameters) is referred to as a validation set. In this chapter, we avoid calling this dataset as a validation set and call it a tuning dataset because it contradicts with the commonly used meaning of validation as “checking the accuracy of” and the definition of validation in 21 CFR 820.3⁷ as “confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use can be consistently fulfilled.” Since cross-validation estimates described above are typically used to modify the trained algorithm, they do not pertain to the finalized production version of the ML algorithm.

Test data are defined as the set of patient-related attributes that are used for characterizing the performance of an ML algorithm and performing appropriate statistical tests. For imaging ML software, the performance is estimated by comparing either the output of the finalized software or the interpretation of a human observer who utilizes the software to a reference standard for each case and summarizing the results for the entire dataset using appropriate metrics.

Collecting a well-characterized and representative dataset is resource-intensive, and therefore, most datasets in medical imaging are much more limited in size, compared to, for example, datasets in natural imaging or electronic health records. A general principle for dataset size is that the training dataset should be large enough to minimize overfitting and the test dataset should be large enough to provide adequate precision in testing, including adequate study power when hypothesis testing is involved. Multiple studies have shown that as the training set is gradually increased starting from a small size, overfitting is initially decreased dramatically, with diminishing returns as the dataset size gets larger [10, 11]. The size for which adding more data provides only diminishing returns depends on the complexity of the ML system and the complexity of the data space. Estimation of the test dataset size for adequate precision and study power is a classical problem in statistics, and pilot data is extremely important for this task.

3.3.2 Independence

A central principle in performance assessment is that the test dataset is required to be independent of the training dataset, meaning that the data for the cases in the test set do not depend on the data for the cases in the training set. It is well-known that the violation of this principle results in optimistically biased performance estimates [12]. To avoid this bias, developers typically set aside a dedicated test data for performance estimation aimed to be independent of the training dataset. There are subtle ways in which the independence principle can be violated if the test dataset is not carefully

⁷ <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=820.3>

selected. We discuss two such mechanisms below. The first is related to including data from a particular patient in both the training and test datasets. The second is related to performing internal validation instead of using an external validation [13] method.

A basic mechanism that can cause a dependence between the training and test sets is the inclusion of data from one patient in both datasets. This could happen if different regions of interest, different image slices, or different objects from the same patients span both the training and test datasets. Since portions of the data from the same patient are expected to be correlated, this practice will result in a statistical dependence between the training and test datasets. A straightforward principle to be followed is to include each patient's data exclusively in the training set or exclusively in the test set.

A more subtle mechanism that can cause a dependence between the training and test datasets is the way the data are sampled or the way that one dataset is partitioned into training and test datasets. Internal validation, which involves partitioning a previously collected sample into training and test datasets randomly or in a stratified way across a given attribute, may result in a dependence between the training and test datasets. Any sampled data, even if it was designed to be collected in a random manner, may not perfectly follow the true distribution of the target population due to finite sample size effects. In addition, there may be a systematic deviation in the feature distribution of a particular sample from the true distribution due to the fact that, for example, the sample may be collected only at a particular site or using only a particular or predominant image acquisition system that does not represent the true distribution. When such a dataset is shuffled and randomly partitioned into training and test datasets, knowledge about the distribution of the training data may provide unfair information about the distribution of the test dataset that would have been impossible to know had the training and test datasets been sampled independently from the true population. A practical approach to reduce this type of dependence is to sample the training and test datasets from multiple different, independent sites, a practice known as external validation [13].

3.3.3 Representativeness

ML algorithms are data-driven, and the distributions of the training and test data have direct implications for algorithm performance and its measurement. Ideally, training and test sets should be large and representative enough so that the collected data provides a good approximation to the true distribution in the target population. As discussed above, well-characterized and annotated medical imaging datasets are typically limited in size. When the dataset size is a constraint, informativeness of a case to be selected for training

the ML algorithm for the task at hand is an additional consideration besides representativeness [14]. Active machine learning techniques aim to proactively select training cases that can best improve model performance, based on informativeness, representativeness, or a combination of the two [15]. Active learning techniques have been applied to train ML algorithms applied to brain imaging [16, 17].

Representativeness of the test dataset is typically desirable when an unbiased estimate of the ML algorithm performance assessment is sought for the target population. For most classification problems, representativeness within each class may be sufficient, which allows designers to enrich the test datasets with classes that have smaller prevalence in the target population. For studies that aim to compare two competing arms (e.g., clinicians' image interpretation with and without ML), enrichment methods that are based on a measurement (e.g., patient or lesion characteristics, risk factors), which trade the unbiased absolute performance results for the practical ability to compare the two competing arms with possible moderate biases, are often acceptable [8]. For example, if cases that are known to be trivial to classify (or diagnose) in both arms of a comparative study are excluded from the test dataset, this will result in a bias in the absolute performance estimates for both arms but may not result in a bias in the difference or change the ranking order of the two arms under comparison, thus allowing the use of a smaller test dataset and a less resource-intensive study design. Likewise, as discussed in Subheading 6.5, when the main goal is to compare the standalone performance of two algorithms to determine which algorithm or modification performs best, it is possible to perform the comparison on a smaller enriched dataset with a careful sampling strategy that does not result in a bias in the difference of the two performance estimates.

3.3.4 Dataset Mismatch

Dataset mismatch is defined as a condition where training and test data follow different distributions, which is popularly known as “dataset shift” in the ML literature [18]. We prefer using “mismatch” because “shift” specifically refers to adding a constant value to each member of a dataset in probability distribution theory, which does not convey all types of mismatches that the term is intended for. Dataset mismatch can also be between test data and real-world deployment data (rather than test and training) or current real-world data vs. future real-world data (e.g., due to changes in clinical practice). There may be many potential reasons for dataset mismatch, with sample selection bias and non-stationary environments cited as the most important ones [19]. Storkey [20] grouped these mismatches into six main categories, including sample selection bias, imbalanced data, simple covariate shift, prior probability shift, domain shift, and source component shift. Dataset

mismatch may result in poor performance of the trained ML algorithm. In addition, especially if caused by a non-stationary environment, dataset mismatch may mean that the performance assessment results obtained at premarket testing may no longer be valid in the clinical environment. A first step in mitigating the effects of dataset mismatch is to detect it. Several methods, including those based on distance measures [21] and dimensionality reduction followed by statistical hypothesis testing [22], have been proposed for this purpose. Techniques for mitigating the effect of dataset mismatch include importance weighting [23] and utilizing stratification, cost curves, or mixture models [24], among others.

3.4 Bias

Bias is a critical factor to consider in study design and analysis for ML assessment, and here we intend to give an overview of sources of bias in ML development and assessment. Note that the general artificial intelligence and machine learning literature currently lacks a consensus on the terminology regarding bias. We consider that performance assessment of an ML system from a finite sample can be cast as a statistical estimation problem. In statistics, a biased estimator is one that provides estimates which are systematically too high or too low [25]. Paralleling this definition, we define statistical bias as a systematic difference between the average performance estimate of an ML system tested in a specified manner and its true performance on the intended population. This systematic difference may result from flaws in any of the components of the assessment framework shown in Fig. 1: collection of patient data and the definition of a reference standard (for both algorithm design and testing stages), algorithm training, analysis methods, and algorithm deployment in the clinic.

Note that the definition of statistical bias above includes systematically different results for different subgroups. ISO/IEC Draft International Standard 22,989 (artificial intelligence concepts and terminology) defines bias as systematic difference in treatment of certain objects, people, or groups in comparison to others, where treatment is any kind of action, including perception, observation, representation, prediction, or decision. As such, statistical bias may result in the type of bias defined in the ISO/IEC Draft International Standard.

We start our discussion of bias with the effect of the dataset representativeness, which has direct implications for ML algorithm performance and its measurement, as described above. When the dataset is not representative of the target population, this can lead to *selection bias*. For example, if all the images in the training or test datasets are acquired with a particular type of scanner while the target patient population may be scanned by many types of scanners, this may lead to an ML algorithm performance estimate that is systematically different from that on the intended population or lead to different results for different subgroups. *Spectrum bias*,

which can be viewed as a consequence of selection bias, describes a systematic error in performance assessment that occurs when the sample of cases studied does not include a complete spectrum of patient and disease characteristics [26]. *Imperfect reference standard bias* and *verification bias* are two types of biases that are related to the reference standard (Subheading 5); the former applies to conditions in which the reference standard procedure is not 100% accurate, and the latter applies to conditions in which only subjects verified for presence or absence of the condition of interest by the reference standard are included in the training set or test set.

Aggregation bias and model design bias are two types of biases that can occur in the algorithm training stage. *Aggregation bias* is related to the information loss which occurs in the substitution of aggregate, or macro-level, data for micro-level data. Aggregation bias can lead to a model that is not optimal for any group or a model that is fit to the dominant population [27]. In ML architecture selection and algorithm training, the designer often has options for model design that may affect the objectives of accuracy, robustness, and fairness, and these objectives may have intrinsic trade-offs. *Model design bias* refers to the design choices that may amplify performance disparities among minority and majority data subgroups [28].

In addition to biases stemming from test dataset composition and the reference standard discussed above, inappropriate selection of the performance metric in the data analysis stage may result in a bias. Many metrics used for evaluation of image analysis algorithms, such as the mean squared error (MSE) for image noise reduction, do not represent the task that the ML algorithm was designed for, e.g., the detection of low-contrast objects in a noisy image. The use of an inappropriate metric may thus result in a difference between the test and true performance, e.g., a conclusion that the algorithm is helpful for its intended use when in clinical reality it is not.

Several factors may contribute to bias after a medical ML system is introduced into the clinic. One of these is the *bias due to a temporal dataset shift* [29] that may cause a mismatch between the data distribution on which the system was developed/tested and the distribution to which the system is applied. Another type of bias, sometimes termed *deployment bias* [27], may be caused by the use of a device in a manner that was not tested as part of the performance assessment and hence does not conform with the intended use of the device, e.g., off-label use. Other types of biases during deployment are also possible because of the differences in the test and clinical environments and unanticipated issues in the integration of the ML system into clinical practice.

4 Algorithm Documentation

4.1 Why Algorithm Documentation Is Important

Machine learning (ML) algorithms have been evolving from traditional techniques with hand-crafted features and interpretable statistical learning models to the more recent deep learning-based neural network models with drastically increased complexity. Appropriate documentation of ML algorithms is critically important for reproducibility and transparency from a scientific point of view. Algorithm description with sufficient details is particularly important in a regulatory setting reviewed by regulators for the assessment of technical quality, for comparing with a legally marketed device, and for the assessment of changes of the algorithm in future versions.

Reproducibility is a well-known cornerstone of science; for scientific findings to be valid and reliable, it is fundamentally important that the experimental procedure is reproducible, whether the experiments are conducted physically or *in silico*. ML studies for detection, diagnosis, or other means of characterization of brain disorders or other diseases are *in silico* experiments involving complex algorithms and big data. As such, we adopt the definition of reproducibility from a National Academies of Sciences, Engineering, and Medicine report [30] as “obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis.” It has been widely recognized that poor documentation such as incomplete data annotation or specification of data processing and analysis is a primary culprit for poor reproducibility in many biomedical studies [31]. Lack of reproducibility may result in not only inconvenience or inferior quality but sometimes a flawed model that can bring real danger to patients when such models are used to tailor treatments in drug clinical trials, as reported by Baggerly and Coombes [32] in their forensic bioinformatics study on a model of gene expression signatures to predict patient response to multidrug regimens.

Appropriate documentation of algorithm design and development is essential for the assessment of technical quality. Identification of the various sources of bias discussed in Subheading 3.4 may not be possible without appropriate algorithm documentation. Furthermore, while there is currently no principled guidance on the design of deep neural network architectures, consensus on good practices and empirical evidence do provide basis for the assessment of technical soundness of an ML algorithm. For example, the choice of loss function is closely related to the clinical task: mean squared error is appropriate for quantification tasks, cross-entropy is often used for classification tasks, and so on. Moreover, the design and optimization of algorithms involve trial-and-error and ad hoc procedures to tune parameters; as such, a developer may introduce bias even unconsciously if the use of patient data and truth labels is not properly documented.

Documentation of ML algorithms is often necessary in a regulatory setting and generally required by FDA under 21 CFR 820.30 design controls. As mentioned in Subheading 2.1, comparison of technological characteristics of a premarket device with a legally marketed predicate device is one of the essential components in determining substantial equivalence for a 510(k) submission. Moreover, the ML algorithm in an FDA-authorized device is often updated, and appropriate algorithm documentation is crucial to decide if a new version has undergone major updates that would require a re-submission to the FDA.

4.2 Essential Elements in Algorithm Description

Many efforts in academia have been devoted to developing checklists for ML algorithm development and reporting to enhance transparency, improve quality, and facilitate reproducibility. A report from the NeurIPS 2019 Reproducibility Program [33] provided a checklist for general machine learning research. Norgeot et al. [34] presented the minimum information about clinical artificial intelligence modeling (MI-CLAIM) checklist as a tool to improve transparency reporting of AI algorithms in medicine. The journal *Radiology* published an editorial with a checklist for artificial intelligence in medical imaging (CLAIM) [35] as a guide for authors and reviewers. Similarly, an editorial of the journal *Medical Physics* introduced a required checklist to ensure rigorous and reproducible research of AI/ML in the field of medical physics [36]. Consensus groups also published the SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials—Artificial Intelligence) as guidelines for clinical trial protocols for interventions involving artificial intelligence [37]. Also, there are undergoing efforts on guidelines for diagnostic and predictive AI models such as the TRIPOD-ML (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis—Machine Learning) [38] and STARD-AI (Standards for Reporting of Diagnostic Accuracy Studies—Artificial Intelligence).

Besides the abovementioned references, the FDA has published a guidance document for premarket notification [510(k)] submissions on computer-assisted detection devices applied to radiology images and radiology device data [39]. Here we provide a list of key elements in describing ML algorithms for medical imaging applications, which we believe are essential (but not necessarily complete) for understanding and technical assessment of an ML algorithm.

- Input.
The types of data the algorithm takes as input may include images and possibly non-imaging data. For input data, essential information includes modality (e.g., CT, MRI, clinical data), compatible acquisition systems (e.g., image scanner manufacturer and model), acquisition parameter ranges (e.g., kVp range, slice thickness in CT imaging), and clinical data collection protocols (e.g., use of contrast agent, MRI sequence).
- Preprocessing.
Input images are often preprocessed such that they are in a suitable form or orientation for further processing. Preprocessing often includes data normalization, which refers to calibration or transformation of image data to that of a reference image (e.g., warping to the reference frame) or to certain numerical range, e.g., slice thickness normalization. Other examples of preprocessing include elimination of irrelevant structures such as a head holder, image size normalization, image orientation normalization, and so on. Sometimes an image quality checker is applied to exclude data with severe artifacts or insufficient quality from further processing and analyses. It is important to describe the specific techniques for normalization and image quality checking. Furthermore, it is critical to make clear how cases failing the quality check are handled clinically (e.g., re-imaging or reviewed by a physician) and account for the excluded cases in the performance assessment.
- Algorithm architecture.
Algorithm architecture is the core module of a machine learning algorithm. In traditional ML techniques, hand-crafted features that are often motivated by physician's experiences are first derived from medical images. A feature selection procedure can be applied to the initially extracted features to select the most useful features for the clinical task of interest. The selected features are then merged by a classifier into a decision variable. There are many choices of the classifier depending on the nature of the data and the purpose of the classifier: linear or quadratic discriminant analysis, k nearest neighbor (k NN) classifiers, artificial neural networks (ANNs), support vector machines, random forests, etc. As such, the algorithm description typically includes the definition of features, the feature selection methods, and the specific classification model. Moreover, it is important to document hyperparameters and the method with which these hyperparameters are determined, for example, the number of neighbors in the k NN method, the number of layers in ANNs, etc.

Recently, deep learning neural network algorithms have been widely used in medical imaging applications. The NN architecture in this type of ML algorithms is composed of a large number of layers that learn to represent data at multiple levels of abstraction and automatically learns features from raw medical image data. As such, instead of sequential hand-crafted feature extraction, selection, and classification, automatic feature engineering and classification (or other types of decision-making such as quantification) are seamlessly integrated in one deep NN architecture. If a published architecture such as AlexNet, VGGNet, Inception V3, etc. is followed exactly, a succinct description is to refer to the reference. Otherwise, the architecture is typically described using a diagram with details such as the number and type of layers, the number of nodes in each layer, the activation functions, the loss function, and so on.

Sometimes hand-crafted features are combined with CNN-based automatic features by a traditional classifier (e.g., random forest) to take advantage of both the power of deep learning in information extraction and domain-specific expertise. In this situation, architecture description includes the entire pipeline, both types of information as described in the above two paragraphs.

- **Algorithm Training.**

ML algorithm training is the process of designing ML algorithm architecture, optimizing the parameters, and selecting the hyperparameters. Taking the popular deep neural networks as an example, the first step in training is to design an architecture or adopt one that has been proven successful in similar applications (see previous bullet). Parameters mainly refer to network weight and bias parameters for combining node outputs in one layer as inputs to nodes of the next layer. Hyperparameters include both those related to network architecture and those related to parameter optimization strategies. Network architecture hyperparameters such as number of hidden layers and units can be pre-selected and fixed if an established architecture is adopted and/or further tuned during training. Another architecture hyperparameter that has been popularly used to avoid overtraining is dropout rate, which refers to the probability of a neuron being “dropped out” in a training step (i.e., the weights are not updated) but may be active in the next step. Hyperparameters related to parameter optimization include learning rate, momentum, number of epochs, batch size, etc.

Given a set of hyperparameters, the network parameters are optimized using training images and associated truth labels. The hyperparameters are typically tuned using a separate tuning dataset. See Subheading 3 for discussion of training data. Again, it is important to fully describe the training process and training data as part of the algorithm description.

- Post-processing and Output.

Given the output of the main ML algorithm, some post-processing steps may be followed, for example, to transform the output to a more interpretable form. The final outputs of an ML algorithm are those that are presented to the end user such as a radiologist or other clinicians. They can be marks on the images indicating the algorithm-determined suspicious areas, a quantitative score indicating the algorithm-estimated likelihood of disease severity, and/or a binary classification indicating if the lesion is benign or malignant, etc. The algorithm description must make clear the final algorithm outputs and how they are intended to be used clinically so that appropriate validation and testing studies can be conducted.

Finally, it should be emphasized that a great description of ML algorithms not only provides these essential elements but also, more usefully, provides rationale on the algorithmic choices. Such rationale may include established good machine learning practices, evidence from similar applications, or methodological research that helps avoid overfitting, reduce bias, and improve generalizability.

5 Reference Standard

Rigorously developed, well-accepted reference standards (also called the “gold standard” or “ground truth”) for training and evaluating machine learning algorithms are essential to validating and characterizing the performance of machine learning algorithms. The reference standard provides a definitive or quasi-definitive characterization of the case based on information that may not be part of the machine learning input, such as biopsy or 1-year follow-up for radiological imaging oncology applications (for an example in the regulatory setting, see⁸). The “truthting” procedures for the cases included in validation (especially external validation) should utilize the best reference standard as recognized by the scientific community to help ensure that the performance of the device is well-characterized. The truthting process is distinct from other aspects of evaluating ML performance as the goal is to determine the “correct” characterization of each case, not to evaluate the device and reader performance in assessing a particular case.

Brain disorders often represent unique challenges to establishing appropriate reference standards. Generally, reference standard can be based on established clinical determination (including an independent modality recognized as a gold standard), follow-up clinical examination, or follow-up medical examination other than imaging. For brain disorders, the pathophysiology may be poorly

⁸ https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN170022.pdf

understood, the progression of some disease may be slow making it difficult to reliably observe changes over time, the clinical definition of the condition may rely heavily on subjective assessments, or definitive assessments may be delayed by years (e.g., Alzheimer's disease) with different syndromes mistaken for the condition of interest (e.g., Parkinsonian-like disorders). In other words, for some brain conditions, the current best available reference standard is based on clinical determination, and confirmation from an alternative method (for instance, histopathological confirmation) may be desirable. Furthermore, ethical and pragmatic challenges of obtaining neurological tissue samples that would allow for independent pathological assessment may limit the utility of biopsy or tissue resection in many brain disorders.

In limited instances, alternatives to independent confirmation of the case "truth," such as interpretation by a reviewing clinician (s), may be considered. Especially in brain disorders where the diagnostic criteria may already be challenging, the importance of multiple reviewing clinicians using the best possible information, even if that requires long-term follow-up, cannot be understated. For some brain disorders such as chronic traumatic encephalopathy or Alzheimer's disease, outstanding challenges remain as the reference standard may be best assessed by biopsy or following death (i.e., autopsy). Greater biological and physiological understanding may be needed to inform the correct diagnosis early in disorder development. Using machine learning techniques to assist in this process is tempting, but the performance will generally be limited by the correctness of the reference standard. In other words, how would we assess if the ML device is outperforming the reference standard as any disagreement may be considered incorrect based on the reference truth?

Uncertainty in the reference standard needs to be accounted for in the analysis. For some machine learning devices, reference standard by expert assessment can be considered, depending on the indications for use, intended use, benefit-risk profile, and device outputs. This is often the case of ML algorithms used in segmentation tasks. In these limited instances, the reference standard from a single clinical truther remains undesirable due to potential concerns about bias or the overall performance of the truther (that is, they are not likely to be 100% accurate, especially for challenging cases). Therefore, multiple clinical truthers are desired. Truthing processes using top experts or truthing processes that weight the clinical truthers' "accuracy" in the construction of the reference standards may also be considered (e.g., see Warfield et al. [40]).

When the truthing process involves interpretation by a reviewing clinician, the number of truthers; their qualifications, experience, and expertise; the instructions for the truthing process; and any other information should be described and documented. In instances where multiple truthers are involved, developer must

consider in advance how the interpretations of these various readers will be incorporated into the final study design and analysis. While combining the interpretation of all truthers into a single reference truth for a particular case may be appropriate in some instances, in other cases such as when the variability between truthers is high, study designs and data analysis methods that take into consideration variability in the reference standard may be appropriate. For instance, reference standard by panel discussions may face unique challenges, especially when loud voices, biases, and group dynamics may influence the outcome. On the other hand, majority vote may lead to other biases such as in segmentation where only including voxels from the majority could lead to small areas or volumes even when compared to all of the participating truthers.

Certain practices in development of the reference standard should be avoided. Often developers look for reference standards of convenience such as a single truther observing the same input data, such as a CT image, as the machine learning algorithm inputs and providing their best judgment as the underlying “truth” of the case. Truthers should not be used as readers who read those images as part of the evaluation of device performance because that can introduce considerable bias to the study results. Public data with unclear processes for establishing the reference standard or incomplete case-level data (such as data without follow-up information, without other typically assessed test results, or incomplete demographic information) frequently raises concerns about the appropriateness of the reference standard in these instances.

The reference standard should generally be based on the best available evidence for the case as recognized by the scientific community. The goal of the reference standard is to establish the “truth” for the outcome of the case. This may present challenges to cohorts where the amount of evidence may differ between cases. Requirements for the minimal amount of information available for a particular case to establish a reasonable “truth” should be defined in advance in the premarket and postmarket setting. As with overall device classification, expectations for rigor and certainty in the reference standard may increase with the device risk associated with misclassification or misdiagnosis. In a regulatory context, often more flexibility is generally permitted in the reference standard for the training data as compared to expectations of rigor in the reference standard for the validation data. Finally, the use of synthetic data is attractive as these techniques provide some opportunities for more well-characterized reference standards in some applications. While synthetic data presents an intriguing approach to addressing some challenges related to reference standards in brain disorders, this is a fairly new topic without significant experience within the current regulatory framework.

6 Standalone Performance Assessment

ML standalone performance is a measure of algorithm performance independent of any human interaction with the ML tool [41]. Standalone performance is the primary assessment for autonomous ML tools that make decisions without clinicians' interactions but may be only one element of assessment for an ML algorithm used as an aid, in which case a clinical assessment of reader performance utilizing the ML may also be required, as discussed in Subheading 7. Standalone testing is also used heavily during algorithm development to benchmark performance and compare potential algorithm modifications before a "final" version is determined. This is because it is often straightforward to integrate iterative testing within the development framework. Standalone testing spans a wide range of possible implementations from initial validation of modifications using a small dataset through large-scale evaluations across multiple independent sites [42] which provides a higher level of confidence in algorithm performance.

Sometimes researchers assume that standalone testing is not important, or at least not as important, as a clinical evaluation, especially for ML-assist devices. However, standalone testing is critical even when a clinical reader study is performed because it is often conducted on larger and more diverse datasets allowing for more refined subgroup analyses and understanding of performance characteristics. It is also critical for assessing the robustness of an ML algorithm and for comparing performance across different algorithms.

In the following, we describe study design, study endpoints, and approaches for assessing standalone performance for specific types of ML tools.

6.1 Segmentation Assessment

Accurate segmentation of brain structures is routinely used in many neurological diseases and conditions when imaging with modalities such as CT, MRI, and PET. As an example, quantitative analysis of brain MRI has been used in assessing brain disorders such as Alzheimer's disease, epilepsy, schizophrenia, multiple sclerosis (MS), cancer, and infectious and degenerative diseases [43]. Often brain assessment quantifying change over time requires the segmentation of brain tissue or anatomy. We define segmentation as the process of partitioning a brain image or image volume into multiple objects defined by a set of voxels unique to each structure or object of interest.

There have been various methods proposed for assessing how well an ML algorithm characterizes objects and how one segmentation algorithm compares to another. Zhang discusses three basic approaches to assessing segmentation algorithms in general [44]. This includes analytical methods, goodness methods, and

discrepancy methods [45]. Analytical methods consider the principles, requirements, utilities, and complexity of segmentation algorithms but can be quite difficult to apply, especially to DL-based segmentation because not all algorithm properties are easily obtained. Goodness methods evaluate segmentation performance by judging the segmented images based on certain quality measures established according to human intuition and include measures such as inter-region uniformity, inter-region contrast, and region shape [45]. Discrepancy methods quantify the difference between segmented objects and a reference standard segmentation. They are the most common type for assessing segmentation algorithms with the caveat that often a ground-truth segmentation is not available. In this case, algorithm segmentations are then compared to human segmentations where the human segmentation is considered the reference standard. Since human segmentations of brain anatomy and structure can be quite variable, segmentations by multiple truthing readers are often collected, and an aggregated reference [40, 46] is used, or the agreement or interchangeability of the algorithm with a truthing reader is assessed.

The remainder of this subsection describes a few common segmentation metrics where we assume a hard segmentation and a single reference standard. A hard segmentation means a voxel is either part of the segmentation or not (this is in contrast to a soft or fuzzy segmentation which means that each voxel is assigned a probability of being part of the segmentation).

An example of a 2D segmentation S and reference segmentation R is shown in Fig. 2 for image X . The false-positive (FP), true-positive (TP), false-negative (FN), and true-negative (TN) regions are also shown. Taha and Hanbury provide a nice overview of 20 segmentation metrics used for discrepancy assessment [47]. Please refer to this paper for more details on many segmentation assessment approaches including methods for assessing fuzzy segmentation algorithms [47]. We next discuss some of the discrepancy assessment approaches frequently used in the literature.

Overlap indexes assess a segmentation by how well it overlaps with the reference. We define some basic overlap metrics below using TP, TN, FP, and FN as voxel counts in the definitions.

- Voxel true-positive rate (TPR), sensitivity, recall: proportion of correctly segmented reference voxels.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- Voxel true-negative rate (TNR), specificity: proportion of correctly segmented background voxels.

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

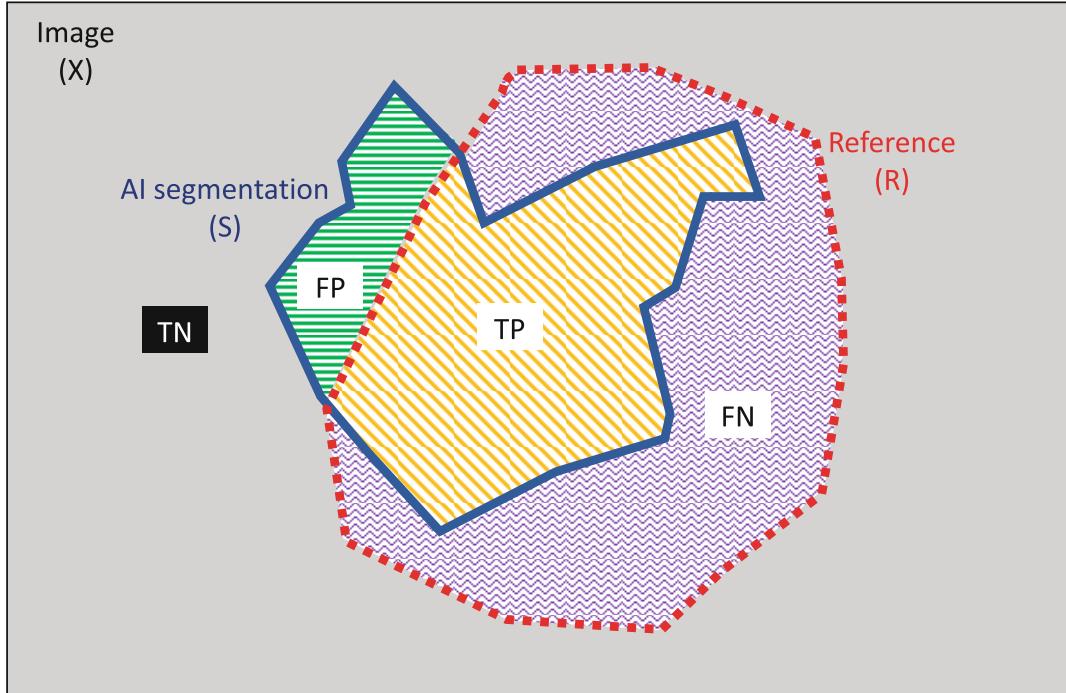


Fig. 2 Diagram of a segmented object (blue solid line) overlaid by the reference segmentation (red dashed line). The false-positive (FP) voxels (green hashed region), true-positive (TP) voxels (yellow hashed region), false-negative (FN) voxels (purple wave region), and true-negative (TN) voxels (gray region) are shown in the figure as well

- Voxel accuracy [45]: proportion of correctly segmented voxels (including both reference and background voxels).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- Dice similarity coefficient (DSC), F1 metric [48]

$$\text{DSC} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} = \frac{2\text{JI}}{1 + \text{JI}}$$

- Jaccard index (JI), intersection over union (IoU) metric [49]

$$\text{JI} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} = \frac{\text{DSC}}{2 - \text{DSC}}$$

The Dice coefficient (DSC) is the most widely used performance metric for characterizing medical image volume segmentations including brain segmentations and can also be used to assess the reproducibility of multiple annotations [47]. The Jaccard index is another common assessment metric. JI and DSC are monotonically related with DSC always having a larger value than JI except at 0 and 1 when the two are equal. However, they have different

properties when averaging performance across multiple segmentations where Jaccard penalizes large segmentation errors more than Dice (somewhat similar to how an L2 norm penalizes larger error more than an L1 norm) [50].

Accuracy is another commonly reported metric, but accuracy is often dominated by a large disparity in the number of reference and background voxels within an image. Accuracy can be high even for poor overlap in the segmentation and reference when the vast majority of voxels in the image are background. This can make it difficult to differentiate between algorithms based only on accuracy differences. A similar observation can be made for specificity or, more generally, for any segmentation metrics involving TN. Indeed, since most voxels are background, TN can be very large. Finally, the definition of the background is not always straightforward and can sometimes be arbitrary (for instance, if the background depends on the field of view of the image).

Distance-based metrics are useful when the boundary of the segmentation is critical [51]. They assess the distance between the segmentation boundary and the reference boundary taking into account the spatial position of the boundary voxels [47]. Some common distance metrics include:

- *Hausdorff distance (HD)* between two voxel sets B_S and B_R (sets of boundary voxels) [47]

$$HD = \max (h(B_S, B_R), h(B_R, B_S)), \text{ where } h(B_R, B_S) = \max_{r \in B_R} \min_{s \in B_S} \|r - s\|$$

- *Mahalanobis distance (MHD)* between two voxel sets B_S and B_R .

$$MHD = \sqrt{(\mu_{B_S} - \mu_{B_R}) S^{-1} (\mu_{B_S} - \mu_{B_R})}, \text{ where } \mu_{B_S} \text{ and } \mu_{B_R} \text{ are the means of the point sets and } S \text{ is the common covariance matrix of the two sets [47]}$$

There are additional segmentation assessment metrics including volume metrics, information theoretic metrics (e.g., mutual information), probabilistic metrics (e.g., intraclass correlation coefficient [ICC]), and pair counting metrics that can also be used to assess the quality of a segmentation algorithm or for comparing multiple segmentations [47].

6.2 Classification Assessment

Classification ML are algorithms designed to parse brain images and data into unique categories. Often the task is differentiating two groups (e.g., cancer versus non-cancer patients), but classification can also be multiclass (e.g., differentiating astrocytoma, glioblastoma, and meningioma brain tumors). The outputs of an ML algorithm can be discrete classes (e.g., via decision tree) or a continuous or a quasi-continuous score (e.g., output of a linear classifier and many DL methods) for an image. As with all ML, the classifier output needs to be assessed and properly interpreted, so

ML performance is understood in the correct context. Tharwat [52] and Hossin and Sulaiman [53] have nice summaries of classification analysis methods. They discuss various performance metrics along with information on how and when each metric might be most effectively used in classifier assessment.

In the remainder of this subsection, we concentrate on binary classifier assessment that includes a wide range of statistical metrics for assessing classifier performance starting with operating point metrics defined directly from discrete ML outputs and moving to more complex metrics based on thresholding a continuous ML output score.

Some basic prevalence-independent metrics (i.e., metrics that do not depend on the prevalence of diseased cases in the standalone database) are described below where TP, TN, FP, and FN are case counts here.

- True-positive rate (TPR), sensitivity, recall

$$\text{TPR} = Se = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- True-negative rate (TNR), specificity

$$\text{TNR} = Sp = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Likelihood ratios are aggregate measures combining sensitivity and specificity. The positive/negative likelihood ratio is the ratio of the probability of a person who has the disease testing positive/negative over the probability of a person who does not have the disease testing positive/negative. They are defined as:

- Positive likelihood ratio (LR^+)

$$LR^+ = \frac{\text{TPR}}{1 - \text{TNR}}$$

- Negative likelihood ratio (LR^-)

$$LR^- = \frac{1 - \text{TPR}}{\text{TNR}}$$

Other operating point metrics depend on the prevalence of disease in the test dataset. They include:

- Positive prediction value (PPV), precision

$$\text{PPV} = \text{Precision} = \frac{\text{TP}}{\text{FP} + \text{TP}}$$

- Negative prediction value (NPV)

$$\text{NPV} = \frac{\text{TN}}{\text{FN} + \text{TN}}$$

- Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- F_1 score

$$F_1 = 2 \cdot \frac{\text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}}$$

These metrics are most appropriate when assessing ML performance in a dataset representing the true clinical population because of their prevalence dependence. They must be interpreted with caution when applied to enriched datasets especially when extrapolating the estimated classification performance to the clinical environment.

For continuous ML scores where a final classification is based on applying a threshold to the output scores, there are aggregation measures that more completely characterize overall classifier performance for a binary task. A common choice is receiver operating characteristic (ROC) analysis which characterizes performance for all possible operating points of the classifier. An ROC curve plots TPR as a function of the false-positive rate ($\text{FPR} = 1 - \text{TNR}$) when the threshold on the classifier output is varied over the complete range of possible output scores [54–56]. An example of an ROC curve is shown in Fig. 3. The advantage of the ROC curve is it shows the benefit (i.e., TPR) as a function of all possible risk values (i.e., FPR) such that a much more complete understanding of the benefit–risk trade-off at all operating points is provided [52].

To facilitate statistical comparisons and to benchmark performance, summary performance can be estimated from ROC curves with the most popular being the area under the ROC curve (AUC) and the partial AUC (PAUC area under just a portion of the ROC curve) [41]. However, the ROC curve should always be plotted to allow for a visual assessment of an individual algorithm's performance or to facilitate a comparison across algorithms. This allows the trade-off across the full range of the ROC curve to be visualized.

Parametric and nonparametric statistical methods are available to both estimate AUC/PAUC and their uncertainties. These approaches allow for statistical comparisons in performance among multiple ML algorithms. There is substantial literature on statistical method for assessing and comparing ROC performance. A great summary of approaches can be found in a report on ROC by the International Commission on Radiation Units and Measurements (ICRU) [57].

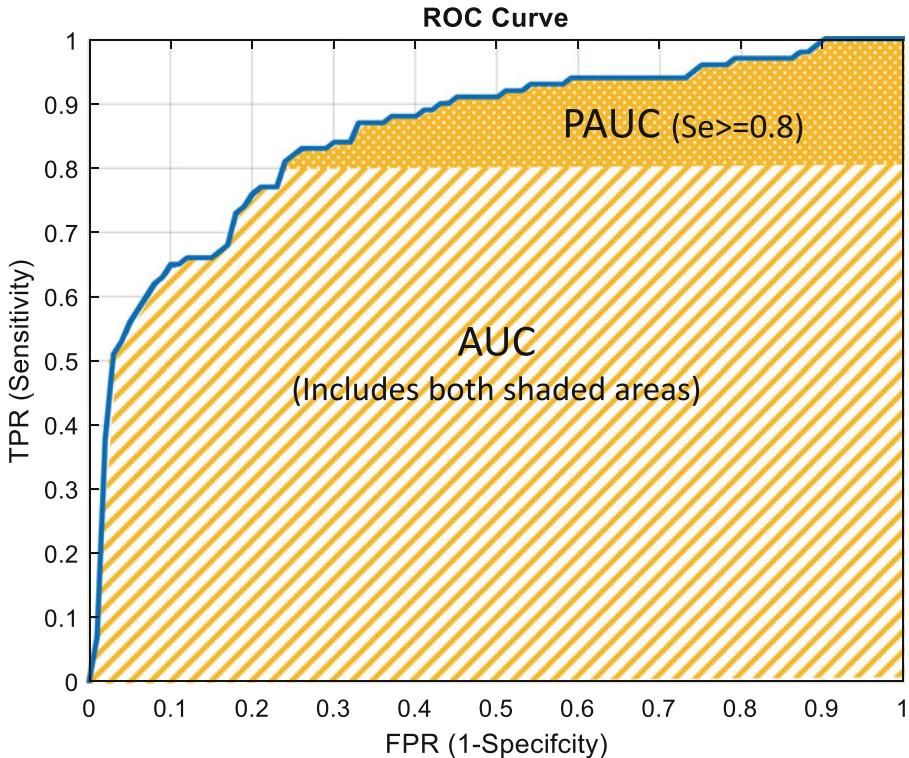


Fig. 3 Plot of an ROC curve for an ML algorithm in a binary classification task. The ROC curve (blue line) shows the trade-off between sensitivity (TPR) and specificity (FPR) for all possible operating points. Both the AUC (includes both shaded regions) and the PAUC for sensitivity ≥ 0.8 are shown in the figure

6.3 Abnormality Detection Assessment

ML detection algorithms mark locations or regions of an image that may reveal abnormalities [41]. Examples of basic ML detection include ML-based bounding boxes or segmentations of potential brain lesions or markers indicating potential brain lesions in an MR or CT scan. Often ML detection outputs include not only localization information but also a confidence score or class determination for the identified regions such that the ML includes both detection and classification functionalities. In the remainder of this subsection, we concentrate on assessing only detection performance without addressing any other potential components of an ML algorithm's output. However, we will still use the ML detection confidence scores, when available, to expand the range of possible performance metrics available for standalone assessment.

Similar to classification metrics, there are a wide range of metrics available for assessing detection performance. Basic detection operating point metrics, usually based on thresholding a continuous ML score for each region, include counts of object-based true-positive (TP), false-positive (FP), and false-negative (FN) detections using the basic definitions in Subheading 6.2 above. Note that object-based true-negative (TN) detections are

generally not estimable in ML detection because there is an infinite (or at least an extremely larger number) of possible TN locations within an image [41]. In addition, ML detection assessment is complicated by the need for a predefined rule (method and threshold) for determining a “correct” detection based on the overlap of a bounding box/segmentation with a reference standard region or the distance from an ML marker to a reference standard object (e.g., distance to the centroid of a reference standard). The overlap metric is often based on the intersection over union (IoU) for bounding boxes/segmentations with a reference standard object and Euclidean distance for markers. However, other potential overlap metrics and criteria may be justifiable for various detection tasks.

Based on the number of TP, FP, and FN detection counts, some basic summary operating point metrics include the true-positive rate (TPR) (i.e., recall) and positive predictive value (PPV) (i.e., precision) that are defined similarly as those in Sub-heading 6.2 but with the unit of regions/cases instead of voxels and the number of FPs per case (or another appropriate unit of interest) since individual cases often include multiple images and abnormalities of interest [41]. For example, an MR exam of the head may include multiple MRI sequences (e.g., T1, T2) such that it is possible to report ML detection performance on a per-patient, per-view (sequence), or per-abnormality (object) basis. The unit of performance should be clearly defined and justified with per-abnormality (or object) performance typically being reported for most image-based ML detection devices especially when only a single exam is available per patient.

Analogous to classification tasks, aggregation metrics that more completely characterize overall ML detection performance are used when a confidence score is available for each detection. ROC analysis is not generally used for ML detection assessment because, as mentioned previously, TNs are not estimable. Therefore, alternate methods have been developed including the free-response receiver operating characteristic (FROC) analysis. FROC accounts for localization and detection of an arbitrary number of abnormalities within an image set [58]. FROC curves plot the fraction of correctly localized lesions as a function of the average number of FPs across the full range of confidence scores for an ML detection algorithm [59]. An example of FROC curve is shown in Fig. 4.

The plot in Fig. 4 shows a nonparametric FROC curve. Parametric FROC methods have been developed using maximum likelihood methods [60–62]. Similar to ROC analysis, FROC area-based metrics can serve as summary performance metrics, but, since the number of FPs in FROC are not bounded, the area under the curve is not limited. This complicates the use of the full area under the FROC curve as a summary figure of merit. Therefore, alternate area-based figures of merit have been developed to summarize and compare FROC performance curves.

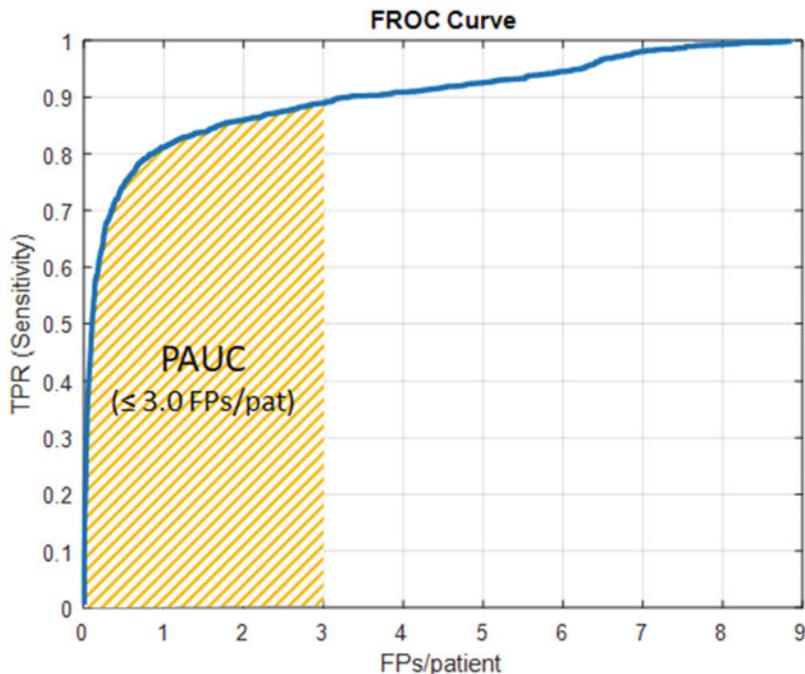


Fig. 4 Plot of an FROC curve for an ML detection algorithm. The FROC curve (blue line) shows the trade-off between object detection sensitivity (TPR) and the number of FPs per patient for all possible operating points. The full area under the FROC curve is not well defined, but a partial area may facilitate comparisons across algorithms. The figure shows a PAUC (shaded region) for ≤ 3.0 FPs/patient. However, AFROC-based summary metrics are more commonly used for characterizing/comparing FROC performance

The area under the alternate FROC (AFROC) curve with a jackknife method (JAFROC) was developed to provide confidence interval estimates and facilitate statistical performance comparisons across algorithms [56, 61]. AFROC provides an alternative way to summarize FROC data where the fraction of negative images falsely called positive are computed based on the highest FP score for each image in the dataset [58]. In this way, the unlimited x-axis of FROC curves is now bounded at 1 as shown in Fig. 5, and the area under the curve is well defined. Chakraborty's jackknife FROC (JAFROC) metric is the area under this AFROC calculated using a jackknife approach [56, 61].

Another common aggregate assessment for ML detection performance is the precision–recall (P–R) curve (*see* Fig. 6) which plots the trade-off between precision and recall across the full range of ML detection algorithm confidence scores [63].

As a reminder, precision (PPV) is a measure of how well the ML detection algorithm identifies only relevant abnormalities, while recall (TPR) is a measure of how well the algorithm finds all abnormalities. A better ML detection algorithm will have a higher precision at a fixed recall. Therefore, a larger area under the P–R

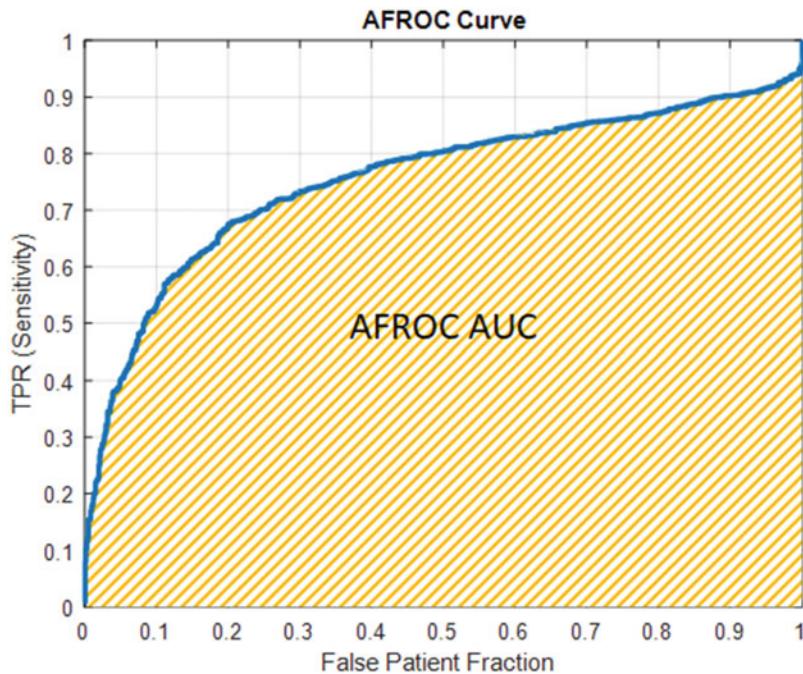


Fig. 5 Plot of an AFROC curve for the same ML detection algorithm given in Fig. 4. The AFROC curve (blue line) shows the trade-off between sensitivity (TPR) and the false patient fraction (fraction of patients with at least one FP) for all possible operating points. The area under the AFROC curve (shaded region) is often used to facilitate comparisons across object detection algorithms

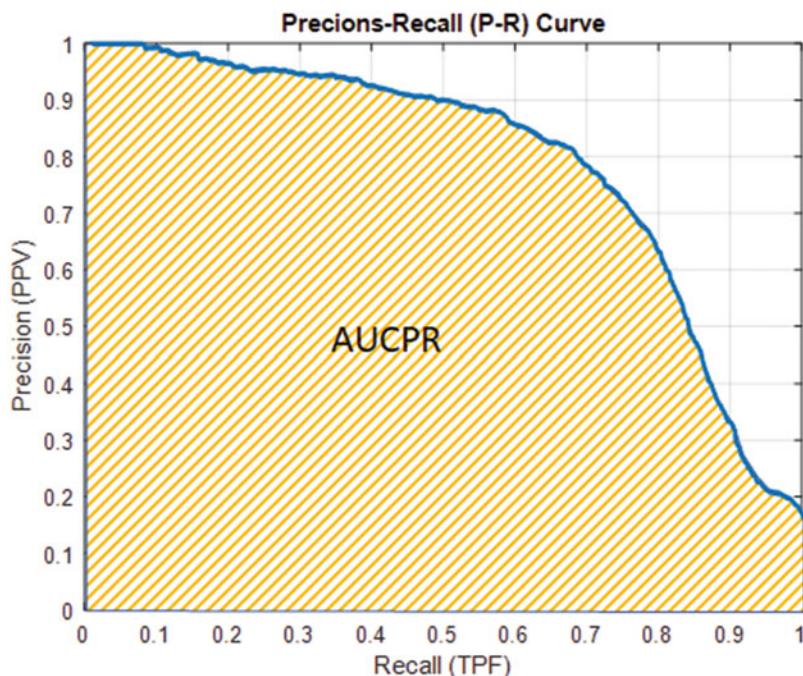


Fig. 6 Plot of a P-R curve for the same ML object detection algorithm as in Figs. 4 and 5. The P-R curve shows the trade-off in precision (PPV) as a function of recall (TPF). The area under the P-R curve (AUCPR) is an aggregate summary metric, for characterizing and comparing P-R curves across object detection algorithms

curve indicates improved performance compared to a competing algorithm, at least when the two P–R curves do not cross. The area under the P–R curve (AUCPR) is again an aggregate summary metric with the average precision (AP) as one estimation method developed in the information retrieval literature and has been used as a performance metric in ML Grand Challenges assessing ML localization algorithms [64, 65]. Nonparametric P–R curve and AP are commonly reported with one definition of AP given below.

- Average precision

$$\begin{aligned} AP &= \sum_{n=1}^N (R_{n+1} - R_n) P_{\text{interp}}(R_{n+1}), \text{ where } P_{\text{interp}}(R_{n+1}) \\ &= \max_{\hat{R}: \hat{R} \geq R_{n+1}} P(\hat{R}) \end{aligned}$$

Another approach is to use an 11-point interpolation by averaging the maximum precision for a set of 11 equally spaced recall levels $[0, 0.1, 0.2, \dots, 1]$ [63]. Parametric [66] and semi-parametric [67] methods for fitting the P–R curve and methods for estimating the AUCPR (e.g., trapezoidal estimators, interpolation estimators) have also been reported in the literature.

One of the complications in assessing an ML algorithm for abnormality detection is the need for determining a “correct” detection based on either an overlap measure for a bounding box/segmentation output or a distance metric for a marker output. Since ML algorithm performance depends on the “correct” detection criterion defined by an empirically chosen overlapping or distance parameter, a sensitivity analysis of the standalone performance across a range of overlap parameters is helpful to confirm that the performance estimate is reasonably stable or to at least understand how the choice of the criterion impacts performance. Moreover, while we have concentrated on detecting a single abnormality here, the abnormality detection metric discussed above can be generalized to multiple-object detection problems by reporting overall performance or assessing performance individually for each type of abnormality and averaging across abnormality types.

6.4 Triage Assessment

A triage ML algorithm analyzes images for findings suggestive of a target clinical condition, but instead of making a diagnosis or detection on the image, the algorithm is limited to generating a notification in the reading worklist or communicating directly to a specialist that a patient has a potential time-sensitive condition. Triage ML devices are often called computer-assisted triage and notification (CADt) devices. CADt is designed to allow a full clinical review earlier in the workflow than without the ML notification, given a true-positive (TP) finding by the algorithm. This can

benefit patients for conditions that are time critical by providing more timely care. For example, in cases involving suspected large vessel occlusion (LVO) stroke, a notification from an effective CADt device could allow a neuro-interventionalist to expeditiously treat the clot, potentially reducing some associated morbidity and mortality.

Another situation that CADt devices are useful is in a busy clinical environment where a large number of cases are queued waiting for clinician review. Instead of reading the cases in a first-in-first-out (FIFO) fashion, the clinician can review CADt flagged cases before non-flagged cases, thereby reducing the waiting time of the diseased patients.

In both situations described above, the sensitivity of the CADt for the target condition is critical so that the truly diseased patients benefit from earlier diagnosis and treatment. However, specificity is also important for the following reasons. An ML algorithm is unlikely to have 100% sensitivity, i.e., there are inevitably false-negative patients in the queue. These patients may be significantly delayed compared with FIFO reading if the triage algorithm has a large false-positive rate (i.e., low specificity). Moreover, too many false alarms may lower the vigilance of a specialist which in turn may affect their performance on the true-positive patients. Therefore, the metric sensitivity and specificity should be used as a pair to assess CADt performance. In the same spirit, the overall capability of the ML algorithm in distinguishing between patients with the condition and those without can be assessed via ROC analysis and the area under the ROC curve.

Despite its usefulness in evaluating a CADt device, the (sensitivity, specificity) pair and ROC performance are metrics of *diagnosis* and, at best, indirect measures of the true clinical effectiveness of an ML triage, i.e., reduction of the waiting time for patients with the target time-sensitive condition. Quantitative assessment of the clinical effectiveness of CADt devices in accelerating the review of patient images with the condition of concern is an open question. Among the efforts we are aware of, Thompson et al. [68] are developing an analytical approach based on the queueing theory to quantify the wait-time-saving of CADt. Under a clinical workflow model parameterized by disease prevalence, patient arrival rate, radiologist service rate, and number of radiologists on-site, their method allows computation of the average waiting time saved for a truly diseased patient due to the use of the CADt device where CADt performance is characterized by its sensitivity and specificity in diagnosing the condition of interest. This approach can potentially be useful in assessing the clinical effectiveness of CADt algorithms but requires further development and validation. Likewise, alternate approaches for assessing true CADt effectiveness in a clinical setting should be an area of continued research.

6.5 Utility of Standalone Performance Assessment

As mentioned previously, ML standalone assessment is primarily used to benchmark algorithm performance and compare with other ML algorithms or prior versions of the same algorithm to determine a performance change. Once a standalone dataset has been established and referenced, and the various performance metrics and criteria set, standalone testing can generally be applied in an efficient manner. Therefore, standalone testing is an important tool for assessing the potential bias in an ML algorithm. When a large diverse standalone dataset containing a range of patients with various demographic characteristics, a wide range of disease conditions, and the full range of acquisition technologies and protocols is available, ML performance can be estimated and compared both overall on the full dataset and in separate subgroups within this larger population to help identify where the ML may perform better and worse.

The standalone testing is also a critical tool for confirming a potential bias or disparity when this disparity is hypothesized, through specifically targeting the assessment to that subgroup of interest. Through standalone testing, ML performance can quickly be evaluated on the specific subgroup to determine if concern is warranted. The data requirements for this type of focused subgroup assessments may not need to be unusually large if the goal is to identify large disparities in performance when the ML algorithm is suspected to be performing poorly. Obviously, identifying more nuanced differences in performance across subgroups requires larger datasets.

Finally, standalone testing is a great tool for comparing ML algorithms. Again, it is ideal to obtain a large diverse real-world dataset to fully assess and benchmark an ML algorithm, but comparison can often be performed on much smaller enriched datasets where the main goal is to determine which algorithm or modification performs best, especially in the developmental phases of an algorithm's life cycle.

6.6 Modifications and Continuous Learning

One of the potential advantages of ML is its ability to quickly learn from new data such that it can remain current to changing patient demographics, clinical practice, and image acquisition technologies. This ability may result in large numbers of updates to an ML algorithm after it becomes available for clinical use. However, each update requires a systematic assessment. Modifications can range from infrequent algorithm updates all the way to continuously learning ML that adapts or learns from real-world experience/data on a continuous basis. This presents a challenge to both ML developers and regulatory bodies such as the FDA.

FDA's traditional paradigm of regulating ML devices is not designed for adaptive technologies, which adapt and optimize performance on a rapid timescale. With this in mind, the FDA is exploring a new, total product life cycle (TPLC) regulatory

approach that may potentially accommodate the rapid modification cycle of ML algorithms allowing for their efficient improvement and adaptation to the changing clinical environment while still providing effective safeguards that meet FDA's statutory requirements to ensure safety and effectiveness. To this end, the FDA released a proposed regulatory framework for modifications to AI/ML software as a medical device (SaMD) in 2019 as a discussion paper [69] requesting feedback from the public on the proposed framework. The proposed TPLC approach is based on [69]:

- The assurance of quality systems and good machine learning practices (GMLP).
- An initial premarket assurance of safety and effectiveness.
- A limited set of SaMD pre-specifications.
- A well-defined algorithm change protocol.

The algorithm change protocol is defined as the specific methods that will be used to achieve and appropriately control the risks of the SaMD pre-specifications [69].

This proposed framework is still under development, but the FDA did provide more details on their potential approach with the release of the AI/ML SaMD Action Plan in January 2021. The Action Plan was developed in response to the stakeholder feedback received on the proposed framework and to support innovative work in the regulation of medical device software and other digital health technologies.⁹

In response to the FDA's proposed framework, Feng et al. have been working to frame an AI/ML algorithm change protocol as an online statistical hypothesis testing problem [70]. The goal of their work was to investigate how "biocreep" resulting from repeated testing and adoption of modifications might lead to a gradual deterioration in ML performance. Feng et al. were able to show that biocreep would regularly occur when using policies with no error-rate guarantees but policies that included error-rate control were able to control biocreep without substantially impacting the ability to approve beneficial modifications [70]. This was an in-depth study of a very limited scope of potential ML modification problems as indicated by Feng et al. [70], and there remains a great deal of work to address the challenges around other types of modifications and conditions. The scientific community, especially interdisciplinary teams of clinicians, statisticians, and domain experts, are encouraged to take on this interesting and complex ML problem [71].

⁹ <https://www.fda.gov/media/106331/download>

7 Clinical Performance Assessment with a Reader Study

Simply put, a reader study for the assessment of ML algorithms is to put the algorithm in the hands of clinicians and study the effectiveness of the algorithm in aiding the clinician's decision-making. In this chapter, a reader study generally refers to a study in which readers (e.g., radiologists) review and interpret medical images for a specified clinical task (e.g., diagnosis) and provide objective quantitative interpretation such as a rating of the likelihood that a condition is present. This is fundamentally different from a survey or questionnaire for the radiologist to indicate if they "like" the functionalities of the ML algorithm, which is not task-specific or particularly subjective (i.e., "beauty test"). Moreover, reader studies for ML in medical imaging typically consist of two arms: reading images without the ML algorithm and with the algorithm output for medical decision-making, thereby enabling a comparison of the reader's performance between with and without the ML aid.

It is fundamentally important to distinguish between *fixed-reader* study and *random-reader* study. When readers are treated as fixed and patient cases are treated as random samples from the patient population, the variability/uncertainty of the performance estimate (without ML or with ML) arises only from the random sample of patient cases. What does this mean? Let us assume we have a radiologist whose name is Barbara in a fixed-reader study and her true diagnostic performance over the entire patient population is A_B . In one experiment, the estimate of Barbara's diagnostic performance is \bar{A}_B with a 95% confidence interval (CI) $[L_{\bar{A}_B}, U_{\bar{A}_B}]$. This means that if the experiment were repeated infinite number of times, *each time with Barbara reading images of a random sample of patients*, then the average of estimates \bar{A}_B in these repeated experiments would be A_B , and the true value A_B would be within the estimated confidence intervals 95% of the time.

In this sense, we say the performance estimate " $\bar{A}_B [L_{\bar{A}_B}, U_{\bar{A}_B}]$ " of *radiologist Barbara* is generalizable to the patient population. Notice that this conclusion is only about Barbara but nobody else.

On the other hand, in a random-reader study where both readers and cases are treated as random effects, the population parameter of interest A is the (average) performance of the reader population over the population of patients. The variability/uncertainty of the performance estimate \bar{A} in one experiment $[L_{\bar{A}}, U_{\bar{A}}]$ should account for both the randomness of readers and that of cases—which is not a trivial task (see next paragraph for relevant literature). The interpretation of such estimates is that, if the experiments were repeated infinite number of times, *each time with a random sample of readers reading a random sample of cases'*

images, then the average of the performance estimates \bar{A} in these repeated experiments would be A , and the population performance A would be within the estimated CIs 95% of the time. In this sense, we say the performance estimate “ $\bar{A} [L_{\bar{A}}, U_{\bar{A}}]$ ” generalizes to both the reader population and the patient population, i.e., the performance estimate represents the expected performance of a random reader reading a random case using a medical device (e.g., an ML algorithm). To distinguish from a fixed-reader study, a random-reader study is often referred to as a multi-reader multi-case (MRMC) study. As a passing note, this discussion also indicates that it is critical to specify the intended patient population and user population of a device so that a study can be designed to collect data from those populations.

The statistical methodology for generalizing the performance of an imaging device to both the population of readers and the population of cases was first developed by Dorfman, Berbaum, and Metz (DBM) [72]. Since then, many methodologies have been developed for the analysis of MRMC data such as the Obuchowski and Rockette (OR) [73] model based on a correlated ANOVA model; the bootstrap method by Beiden, Wagner, and Campbell [74]; and the U statistic method by Gallas [75]. Relationships among these methods have also been investigated [76, 77]. These early developments of MRMC analysis methods have focused on the area under the ROC curve (AUC) as a performance metric; some of these methods (e.g., OR and U statistic methods) have been extended to binary performance metrics [78], and all these methods have been validated with simulation studies [79] [80]. Some of these methods also have publicly available software tools, such as the integrated and updated OR–DBM method¹⁰ and the U statistic method.¹¹

The most widely used MRMC study design for comparing two modalities (e.g., without ML versus with ML) is the fully crossed (FC) design, in which every reader reads every case in both modalities. The advantage of pairing both readers and cases across two modalities is that it builds a positive correlation between the performance estimates of the two modalities, thereby reducing the variability of the performance difference and enhancing the power of detecting the performance difference. This reduction of variability can be easily appreciated by a simple formula

¹⁰ Software | Medical Image Perception Laboratory Department of Radiology (uiowa.edu): <https://perception.lab.uiowa.edu/software-0>

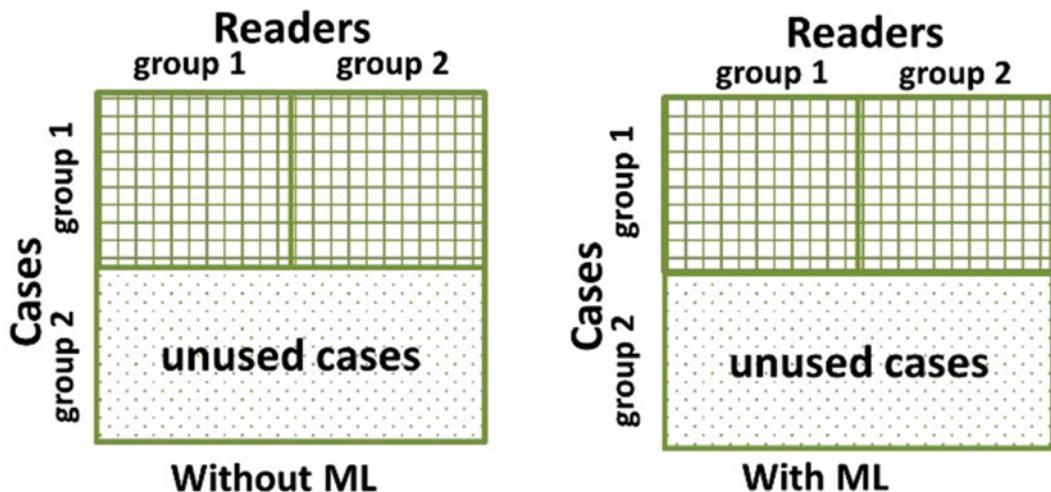
¹¹ iMRMC: Software to do multi-reader multi-case analysis of reader studies: <https://github.com/DIDSR/iMRMC>

$$\text{Var}[\widehat{A}_1 - \widehat{A}_2] = \text{Var}[\widehat{A}_1] + \text{Var}[\widehat{A}_2] - 2\rho\sqrt{\text{Var}[\widehat{A}_1]\text{Var}[\widehat{A}_2]},$$

where “Var” denotes variance; \widehat{A}_1 and \widehat{A}_2 are performance estimates for, e.g., without ML and with ML, respectively; and ρ is the correlation between \widehat{A}_1 and \widehat{A}_2 and is positive under normal circumstances. Pairing cases from two modalities sometimes is not advised due to safety concerns, for example, if both imaging modalities involve ionizing radiation to patients, imaging the patient twice may raise dose concerns. Fortunately, this is not generally an issue for the assessment of ML algorithms, and pairing cases in a “without ML versus with ML” comparison is feasible in many diagnostic situations.

The FC design has been regarded as the most powerful design in the sense that it makes full use of available readers and cases in collection of information. However, practically the workload of a radiologist may be limited, and oftentimes an investigator may have more cases than what readers can afford to read. Moreover, as multi-site evaluation becomes popular for better generalizability, the transfer of cases among different clinical sites can be logically demanding. To overcome these limitations, Obuchowski [81] investigated the split-plot design, where different groups of readers read different groups of cases. The combined reader/case group can still be paired across modalities to reduce the variability of performance difference. Figure 7 provides a visual illustration of the FC design and the paired split-plot (PSP) design. What might be surprising is that the PSP design can be more powerful than the FC design, as shown by Hillis et al. with empirical data [82] and Chen et al. with both theoretical analysis and real-world data [83]. This may sound like a paradox since the FC design is regarded as “the most powerful design,” but it is not. Referring to Fig. 7, suppose we have a certain number of readers and each of them can read the same number of cases. In the FC design, all the readers read the same cases (*see* Fig. 7, top), whereas, in the PSP design, readers are partitioned into two groups with each group reading the same number of cases from two different case sets (*see* Fig. 7, bottom). As such, the two designs involve the same amount of workload (i.e., number of image interpretations). However, the PSP design has reduced variability in performance estimates and performance difference estimates and hence increased statistical power, as proved by Chen et al. [83] because of the inclusion of additional cases. One way to understand this is that, with the same workload, reading difference cases (by half of the readers) gains more information than reading the same cases. This is also consistent with a common statistical sense: when we have more cases, the variability of the “mean” measured on the cases is reduced. In summary, the FC design is the most powerful *given the same*

Fully Crossed Design



Paired Split-Plot Design

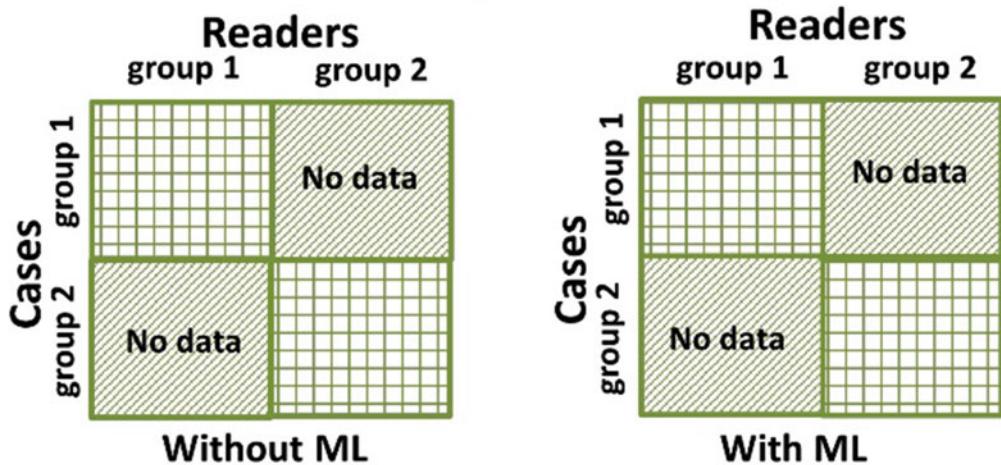


Fig. 7 Illustration of the fully crossed design and the paired split-plot design. The squares with grid can be understood as the data matrix collected in the reader study with each row representing a case, each column representing a reader, and each data element representing the rating of the case by the reader

number of readers and cases, but the PSP design can be more powerful given the same number of image interpretations with a price of collecting extra cases [83].

The design of an MRMC reader study involves a great deal of considerations including patient data collection (see Subheading 3), establishment of a reference standard (see Subheading 5), and many other aspects such as the recruitment and training of readers and

reading session design, e.g., *sequential reading*, where the readers read images with ML turned on immediately after reading without ML, or *concurrent reading*, where readers read images with ML turned on from the very beginning and this is typically compared against readers' performance reading images without ML in a separate session. It is worth noting that the discussion of the performance testing here is generally based on ML systems that are intended to "aid" or interface with an expert radiologist. The intended use of a model may warrant additional testing considerations related to human factors and human interpretability depending on how the model is integrated into the clinical workflow. Moreover, MRMC studies for the assessment of ML in imaging are often retrospective and controlled "laboratory" studies, in which typically only information related to the device of interest is presented to the readers (e.g., "image only" versus "image plus ML output"), whereas in real-world clinical practice, more information is often available to the physician, e.g., patient history, clinical tests, and/or other types of imaging exams. The diseased cases are often enriched when the natural prevalence is low in controlled laboratory studies. The purpose of such designs is to remove certain confounders and increase the statistical power to study the impact of the ML algorithm itself rather than the "absolute" performance of clinicians in the real world (as discussed in Subheading 3). However, consideration should still be given to ensure the study execution is as close to the clinical environment as possible and identify/mitigate potential biases, for example, the readers should be trained to use the ML algorithm as if they were instructed in the clinic. It is also important to randomize cases, readers, and reading sessions to minimize bias. For more details on the design of MRMC studies, interested readers can refer to an FDA guidance document [84], a consensus paper by Gallas et al. [8], as well as a tutorial paper by Wagner et al. [55].

8 Statistical Analysis

The statistical analysis plays a critical role in the assessment of ML performance but may be under-appreciated by many ML developers. For example, there are still publications that present point estimates of ML performance without quantification of uncertainties (standard deviations, confidence intervals). Even if uncertainty estimates are provided, the methods of uncertainty estimation are sometimes unclear or even inappropriate. Another example is the re-use of test data. One may follow the good practice of using independent datasets for ML training and testing. However, if the test data is repeatedly used, the seemingly innocent good practice may introduce optimistic bias to the performance estimate or even lead to a spurious discovery because the repeatedly measured

performance on the test dataset may inform training of the algorithm to adaptively fit the test data [85, 86]. As the quote goes, “if you torture the data long enough, it will confess.” The lesson is, without following appropriate statistical principles, ML developers may be led to a blind alley due to statistical pitfalls: comparisons are made without statistical rigor, conclusions are drawn without appropriate data to substantiate, and spurious findings out of overfitting are celebrated. Statistical practices have a major impact on the ability to conduct reproducible research.

A good practice to avoid such pitfalls is, for any performance assessment study—either standalone performance assessment or an MRMC study—to *pre-specify* a statistical analysis plan (SAP) with valid statistical methods. The word “pre-specify” is emphasized because post hoc analyses can inflate the experiment-wise type I error rate and endanger the scientific validity of an otherwise well-designed and well-conducted study. Below we list exemplar elements in an SAP for ML development and assessment. We note that not all of them are necessarily applicable to a specific study. A specific SAP should be consistent with the study objectives, designs, nature of data, and statistical analysis methods.

1. Primary hypotheses and secondary hypotheses that are consistent with the primary and secondary goals of a study. This also necessarily involves choosing appropriate performance metrics (*see Subheading 6* for different metrics corresponding to different clinical tasks). For example, the primary goal of an MRMC study might be to show the radiologists using an ML algorithm perform significantly better than without using the algorithm in the task of distinguishing between benign and malignant brain tumors, and a secondary goal might be to show the radiologist using an ML algorithm has significantly better specificity (S_p) at a given sensitivity. Then the null (H_0) and alternative (H_1) primary hypotheses can be stated as

$$H_0 : \text{AUC}_{\text{with ML}} = \text{AUC}_{\text{without ML}}; H_1 \\ : \text{AUC}_{\text{with ML}} > \text{AUC}_{\text{without ML}}.$$

And the secondary hypotheses may be stated as

$$H_0 : S_{p\text{with ML}} = S_{p\text{without ML}}; H_1 : S_{p\text{with ML}} > S_{p\text{without ML}}.$$

2. A plan for use of patient data in various stages of ML algorithm development and performance assessment. As discussed in Subheading 3, patient data are used in both the development and assessment of ML algorithms. A pre-specified plan for appropriate use of patient data is crucial for achieving the goals of algorithm development and performance validation and controlling various sources of bias in the process.

3. Methods for analyzing the study data to estimate the pre-specified performance metric, the uncertainty of the metric (e.g., standard deviations and confidence intervals) accounting for all sources of variability including reference standard as needed, and the test statistic for hypothesis testing. It is critically important to examine if the assumptions behind the statistical methods are appropriate for the data and, when necessary, use an alternative method to verify the results.
4. Sample size determination. In a standalone performance assessment study, this is to determine the number of patients to be included in the study such that the study data are representative of the intended patient population (*see* Subheading 3.3.3) and, when applicable, the study has sufficient statistical power (typically set to be >80%) to detect a significant effect (e.g., superior performance compared with a control). With a single source of variability, standard statistical methods and software tools are often useful for sizing a standalone performance assessment study.

In an MRMC study, both the number of readers and the number of cases need to be determined. Sample size determination is again mainly for assuring a reasonable chance of success in the study planning stage. From a technical point of view (i.e., not taking into consideration practical issues such as budget), sample size is typically determined by considering (1) that the sample sizes are large enough to include samples that represent the intended patient and reader populations and (2) the sample sizes are sufficient to achieve a target statistical power in a hypothesis testing study. Due to the complexity, specialized software tools can be used for sizing an MRMC study [87], and the MRMC software tools cited in Subheading 7 provide the sizing functionality.

The information needed for sizing a pivotal study is often obtained in a pilot study, as discussed in Subheading 3. However, sometimes the pilot study is too limited to provide reliable information, and one may find attempting to re-size a pivotal study after an interim analysis of the data. Naively re-sizing the study based on information obtained in the same study may inflate the type I error rate. Huang et al. [88] developed an approach that allows adaptive re-sizing of an MRMC study with information obtained in an interim analysis such that the statistical power is adjusted to a target value and the type I error rate is retained by paying a statistical penalty in the final hypothesis testing.

5. A plan for adjusting p-values and/or confidence intervals for multiple comparisons or hypothesis tests.
6. A plan for handling missing data and assessing the impact of missing data (e.g., missing reader data, missing follow-up data

to confirm negative results) on the study conclusions. Although statistical techniques may be used to address issues of loss-to-follow-up and missing data, these techniques often employ major assumptions that cannot be fully validated for a particular study. Therefore, the best way to address issues of missing data due to loss-to-follow-up is to plan to minimize its occurrence during the planning and management of the clinical study. Nevertheless, the study protocol should pre-specify appropriate statistical data analysis methods, in addition to sensitivity analyses, for handling missing data.

9 Summary Remarks

In this chapter, we provided an overview of a performance assessment framework for imaging-based ML algorithms. We discussed general considerations in study design and data collection, establishment of a reference standard, algorithm documentation, algorithm standalone performance as well as clinical reader studies, and statistical data analysis in performance testing. We believe that these topics are relevant not only in the regulatory setting but also to reproducible science and technology development. Because patient data and clinical experts' annotations are used in both the *development* and *assessment* of ML algorithms, performance assessment should be considered from the very beginning of development to make efficient use of available data. In addition, performance assessment and algorithm development (e.g., tuning, internal validation) are often iterative; meaningful assessment methodologies and tools are not only meant to make the assessment *rigorous* but also *cost-effective*. Furthermore, performance assessment methodologies are also tremendously helpful to assure quality and reproducibility, control bias, and avoid pitfalls and blind alleys.

Machine learning technologies are still rapidly evolving, and their applications in medicine and brain imaging in particular are expanding. It is widely recognized that ML is playing a pivotal role in revolutionizing medicine and promoting public health to a new level. Accompanying these potential developments are new research questions on assessment methodologies. We have touched upon topics in this chapter such as novel types of clinical applications enabled by ML and continuous learning ML. Other exciting topics may include improvement and assessment of robustness and generalizability of ML algorithms, synthetic data augmentation, characterization of bias/fairness, and uncertainty-aware ML algorithms that output not only clinical conditions of interest but also "I don't know," among many others. We believe that assessment methodologies and regulatory science play a critical role in fully realizing the great potential of ML in medicine, in facilitating ML device innovation, and in accelerating the translation of these technologies from bench to bedside to the benefit of patients.

Acknowledgments

The authors thank Drs. Robert Ochs, Alexej Gossmann, and Aldo Badano for carefully reviewing the manuscript and providing helpful comments.

References

1. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, Summers RM, Giger ML (2019) Deep learning in medical imaging and radiation therapy. *Med Phys* 46(1):e1–e36. <https://doi.org/10.1002/mp.13264>
2. Lui YW, Chang PD, Zaharchuk G, Barboriak DP, Flanders AE, Wintermark M, Hess CP, Filippi CG (2020) Artificial intelligence in neuroradiology: current status and future directions. *Am J Neuroradiol* 41(8):E52–E59. <https://doi.org/10.3174/ajnr.A6681>
3. U.S. Food and Drug Administration (2017) De Novo classification process (Evaluation of Automatic Class III Designation). Guidance for Industry and Food and Drug Administration Staff
4. U.S. Food and Drug Administration (2014) The 510(k) program: evaluating substantial equivalence in premarket notifications [510 (k)]. Guidance for Industry and Food and Drug Administration Staff
5. U.S. Food and Drug Administration (2012) Factors to consider when making benefit-risk determinations in medical device premarket approval and De Novo classifications. Guidance for Industry and Food and Drug Administration Staff
6. U.S. Food and Drug Administration (2018) Benefit-risk factors to consider when determining Substantial equivalence in premarket notifications (510(k)) with different technological characteristics. Guidance for Industry and Food and Drug Administration Staff
7. U.S. Food and Drug Administration (2021) Requests for feedback and meetings for medical device submissions: the Q-submission program. Guidance for Industry and Food and Drug Administration Staff
8. Gallas BD, Chan HP, D'Orsi CJ, Dodd LE, Giger ML, Gur D, Krupinski EA, Metz CE, Myers KJ, Obuchowski NA, Sahiner B, Toldano AY, Zuley ML (2012) Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. *Acad Radiol* 19(4):463–477. <https://doi.org/10.1016/j.acra.2011.12.016>
9. Hastie T, Tibshirani R, Friedman J (2017) The elements of statistical learning. Series in statistics, 2nd (corrected 12th printing) edn. Springer, New York
10. Chan H-P, Sahiner B, Wagner RF, Petrick N (1999) Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers. *Med Phys* 26(12):2654–2668
11. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster R (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316(22):2402–2410. <https://doi.org/10.1001/jama.2016.17216>
12. Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic Press, New York
13. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 162(1):W1–W73. <https://doi.org/10.7326/m14-0698>
14. Du B, Wang Z, Zhang L, Zhang L, Liu W, Shen J, Tao D (2019) Exploring representativeness and informativeness for active learning. arXiv:1904.06685
15. Huang S, Jin R, Zhou Z (2014) Active learning by querying informative and representative examples. *IEEE Trans Pattern Anal Mach Intell* 36:1936–1949
16. Sharma D, Shanis Z, Reddy CK, Gerber S, Enquobahrie A (2019) Active learning technique for multimodal brain tumor segmentation using limited labeled images. In: Wang Q, Milletari F, Nguyen HV et al (eds) Domain adaptation and representation transfer and medical image learning with less labels and imperfect data. Springer International Publishing, Cham, pp 148–156

17. Hao R, Namdar K, Liu L, Khalvati F (2021) A transfer learning-based active learning framework for brain tumor classification. *Front Artif Intell* 4(61):635766. <https://doi.org/10.3389/frai.2021.635766>
18. Quionero-Candela J, Sugiyama M, Schwaighofer A, Lawrence N (2009) Dataset shift in machine learning. MIT Press, Cambridge MA
19. Moreno-Torres JG, Raeder T, Alaiz-Rodriguez R, Chawla NV, Herrera F (2012) A unifying view on dataset shift in classification. *Pattern Recogn* 45(1):521–530. <https://doi.org/10.1016/j.patcog.2011.06.019>
20. Storkey A (2009) When training and test sets are different: characterizing learning transfer. In: Quionero-Candela J, Sugiyama M, Schwaighofer A, Lawrence N (eds) *Dataset shift in machine learning*. MIT Press, Cambridge, MA, pp 3–28
21. Goldenberg I, Webb G (2019) Survey of distance measures for quantifying concept drift and shift in numeric data. *Knowl Inf Syst* 60: 591–615. <https://doi.org/10.1007/s10115-018-1257-z>
22. Rabanser S, Günemann S, Lipton ZC (2018) Failing loudly: an empirical study of methods for detecting dataset shift. arXiv:1810.11953
23. Dockès J, Varoquaux G, Poline J-B (2021) Preventing dataset shift from breaking machine-learning biomarkers. arXiv:2107.09947
24. Turhan B (2012) On the dataset shift problem in software engineering prediction models. *Empir Softw Eng* 17(1):62–74. <https://doi.org/10.1007/s10664-011-9182-8>
25. U.S. Food and Drug Administration (2007) Guidance for industry and FDA staff: statistical guidance on reporting results from studies evaluating diagnostic tests. vol 2007. U.S Food and Drug Administration, Silver Spring
26. Zhou XH, Obuchowski NA, McClish DK (2002) Statistical methods in diagnostic medicine. Wiley
27. Suresh H, Guttag JV (2021) A framework for understanding sources of harm throughout the machine learning life cycle. arXiv:190110002 [cs, stat]
28. Hooker S (2021) Moving beyond “algorithmic bias is a data problem”. *Patterns* 2(4):100241. <https://doi.org/10.1016/j.patter.2021.100241>
29. Guo LL, Pfohl SR, Fries J, Posada J, Fleming SL, Aftandilian C, Shah N, Sung L (2021) Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Appl Clin Inform* 12(4):808–815. <https://doi.org/10.1055/s-0041-1735184>
30. National Academies of Sciences E, Medicine (2019) Reproducibility and replicability in science. The National Academies Press, Washington, DC. <https://doi.org/10.17226/25303>
31. Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, van Noort V (2009) Repeatability of published microarray gene expression analyses. *Nat Genet* 41(2):149–155. <https://doi.org/10.1038/ng.295>
32. Baggerly KA, Coombes KR (2009) Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *Ann Appl Stat* 3(4): 1309–1334, 1326
33. Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d’Alché-Buc F, Fox E, Larochelle H (2020) Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). arXiv:2003.12206
34. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, Arnaout R, Kohane IS, Saria S, Topol E, Obermeyer Z, Yu B, Butte AJ (2020) Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* 26(9):1320–1324. <https://doi.org/10.1038/s41591-020-1041-y>
35. Mongan J, Moy L, Charles E, Kahn J (2020) Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell* 2(2):e200029. <https://doi.org/10.1148/ryai.2020200029>
36. El Naqa I, Boone JM, Benedict SH, Goodsitt MM, Chan HP, Drukker K, Hadjiiski L, Ruan D, Sahiner B (2021) AI in medical physics: guidelines for publication. *Med Phys* 48(9):4711–4714. <https://doi.org/10.1002/mp.15170>
37. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ, Darzi A, Holmes C, Yau C, Moher D, Ashrafian H, Deeks JJ, Ferrante di Ruffano L, Faes L, Keane PA, Vollmer SJ, Lee AY, Jonas A, Esteva A, Beam AL, Panico MB, Lee CS, Haug C, Kelly CJ, Yau C, Mulrow C, Espinoza C, Fletcher J, Moher D, Paltoo D, Manna E, Price G, Collins GS, Harvey H, Matcham J, Monteiro J, ElZarrad MK, Ferrante di Ruffano L, Oakden-Rayner L, McCradden M, Keane PA, Savage R, Golub R, Sarkar R, Rowley S, The S-A, Group C-AW, Spirit AI, Group C-AS, Spirit

- AI, Group C-AC (2020) Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* 26(9):1351–1363. <https://doi.org/10.1038/s41591-020-1037-7>
38. Collins G, Moons K (2019) Reporting of artificial intelligence prediction models. *Lancet* 393:1577–1579. [https://doi.org/10.1016/S0140-6736\(19\)30037-6](https://doi.org/10.1016/S0140-6736(19)30037-6)
39. U.S. Food and Drug Administration (2012) Clinical performance assessment: Considerations for computer-assisted detection devices applied to radiology images and radiology device data – premarket approval (PMA) and premarket notification [510(k)] submissions – Guidance for industry and FDA staff. <https://www.fda.gov/media/77642/download>. Accessed 31 Oct 2021
40. Warfield SK, Zou KH, Wells WM (2004) Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 23(7):903–921
41. Petrick N, Sahiner B, Armato SG III, Bert A, Correale L, Delsanto S, Freedman MT, Fryd D, Gur D, Hadjiiski L, Huo Z, Jiang Y, Morra L, Paquerault S, Raykar V, Salganicoff M, Samuelson F, Summers RM, Tournassi G, Yoshida H, Zheng B, Zhou C, Chan H-P (2013) Evaluation of computer-aided detection and diagnosis systems. *Med Phys* 40:087001–087017
42. Steyerberg EW (2019) Overfitting and optimism in prediction models. In: *Clinical prediction models*. Springer, pp 95–112
43. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ (2017) Deep learning for brain MRI segmentation: state of the art and future directions. *J Digit Imaging* 30(4): 449–459. <https://doi.org/10.1007/s10278-017-9983-4>
44. Zhang YJ (1996) A survey on evaluation methods for image segmentation. *Pattern Recogn* 29(8):1335–1346. [https://doi.org/10.1016/0031-3203\(95\)00169-7](https://doi.org/10.1016/0031-3203(95)00169-7)
45. Zhang YJ (2001) A review of recent evaluation methods for image segmentation. In: *Proceedings of the sixth international symposium on signal processing and its applications (Cat. No.01EX467)*, 13–16 Aug 2001. vol. 141, pp 148–151. <https://doi.org/10.1109/ISSPA.2001.949797>
46. Meyer CR, Johnson TD, McLennan G, Aberle DR, Kazerooni EA, Macmahon H, Mullan BF, Yankelevitz DF, van Beek EJR, Armato SG 3rd, McNitt-Gray MF, Reeves AP, Gur D, Henschke CI, Hoffman EA, Bland PH, Laderach G, Pais R, Qing D, Piker C, Guo J, Starkey A, Max D, Croft BY, Clarke LP (2006) Evaluation of lung MDCT nodule annotation across radiologists and methods. *Acad Radiol* 13(10):1254–1265
47. Taha AA, Hanbury A (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging* 15(1):29. <https://doi.org/10.1186/s12880-015-0068-x>
48. Dice LR (1945) Measures of the amount of ecological association between species. *Ecology* 26(3):297–302. <https://doi.org/10.2307/1932409>
49. Jaccard P (1912) The distribution of the flora in the alpine zone. *New Phytol* 11(2):37–50
50. Willem (2017) FI/Dice-Score vs IoU. Cross Validated. <https://stats.stackexchange.com/questions/273537/f1-dice-score-vs-iou/276144#276144>. Accessed 9/29/2021
51. Fenster A, Chiu B (2005) Evaluation of segmentation algorithms for medical imaging. In: *2005 IEEE engineering in medicine and biology 27th annual conference*, 17–18 Jan 2006. pp 7186–7189. <https://doi.org/10.1109/IEMBS.2005.1616166>
52. Tharwat A (2021) Classification assessment methods. *Appl Comput Inform* 17(1): 168–192. <https://doi.org/10.1016/j.aci.2018.08.003>
53. Hossin M, Sulaiman MN (2015) A review on evaluation metrics for data classification evaluations. *Int J Data Min Knowl Process* 5(2):1
54. Obuchowski NA (2003) Receiver operating characteristic curves and their use in radiology. *Radiology* 229(1):3–8
55. Wagner RF, Metz CE, Campbell G (2007) Assessment of medical imaging systems and computer aids: a tutorial review. *Acad Radiol* 14(6):723–748
56. Chakraborty DP (2018) Observer performance methods for diagnostic imaging: foundations, modeling, and applications with r-based examples. *Imaging in medical diagnosis and therapy*. CRC Press, Boca Raton, FL
57. ICRU (2008) Receiver operating characteristic analysis in medical imaging. Report 79. International Commission of Radiation Units and Measurements, Bethesda, MD
58. He X, Frey E (2009) ROC, LROC, FROC, AFROC: an alphabet soup. *J Am Coll Radiol* 6(9):652–655
59. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH (1977) A free response approach to the measurement and characterization of radiographic observer performance. *Proc SPIE* 127: 124–135

60. Edwards DC, Kupinski MA, Metz CE, Nishikawa RM (2002) Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. *Med Phys* 29(12):2861–2870. <https://doi.org/10.1118/1.1524631>
61. Chakraborty DP (2006) Analysis of location specific observer performance data: validated extensions of the jackknife free-response (JAFROC) method. *Acad Radiol* 13(10): 1187–1193
62. Chakraborty DP (2006) A search model and figure of merit for observer data acquired according to the free-response paradigm. *Phys Med Biol* 51(14):3449–3462
63. Padilla R, Netto SL, Silva EABd (2020) A survey on performance metrics for object-detection algorithms. In: 2020 international conference on systems, signals and image processing (IWSSIP), 1–3 July 2020. pp 237–242. <https://doi.org/10.1109/IWSSIP48289.2020.9145130>
64. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The Pascal visual object classes (VOC) challenge. *Int J Comput Vis* 88(2):303–338. <https://doi.org/10.1007/s11263-009-0275-4>
65. ImageNet (2017) ImageNet object localization challenge. Kaggle. <https://www.kaggle.com/c/imagenet-object-localization-challenge/>. Accessed 10/22/2021 2021
66. Liu Z, Bondell HD (2019) Binormal precision-recall curves for optimal classification of imbalanced data. *Stat Biosci* 11(1):141–161. <https://doi.org/10.1007/s12561-019-09231-9>
67. Sahiner B, Chen W, Pezeshk A, Petrick N (2016) Semi-parametric estimation of the area under the precision-recall curve. In: SPIE medical imaging. International Society for Optics and Photonics, pp 97870D-97870D-97877
68. Thompson E, Levine G, Chen W, Sahiner B, Li Q, Petrick N, Samuelson F (2022) Wait-time-saving analysis and clinical effectiveness of computer-aided triage and notification (CADt) devices based on queueing theory. In: Taylor-Phillips CRM-TaS (ed) Medical imaging 2022: Image perception, observer performance, and technology assessment, San Diego, CA, SPIE, p accepted
69. U.S. Food and Drug Administration (2019) Proposed regulatory framework for modifications to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SaMD) – Discussion paper and request for feedback. U.S Food and Drug Administration. <https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf>. Accessed 31 Oct 2021
70. Feng J, Emerson S, Simon N (2021) Approval policies for modifications to machine learning-based software as a medical device: a study of bio-creep. *Biometrics* 77(1):31–44. <https://doi.org/10.1111/biom.13379>
71. Pennello G, Sahiner B, Gossman A, Petrick N (2021) Discussion on “approval policies for modifications to machine learning-based software as a medical device: a study of bio-creep” by Jean Feng, Scott Emerson, and Noah Simon. *Biometrics* 77(1):45–48. <https://doi.org/10.1111/biom.13381>
72. Dorfman DD, Berbaum KS, Metz CE (1992) Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Investig Radiol* 27(9):723–731
73. Obuchowski NA, Rockette HE (1995) Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. *Commun Stat Simul Comput* 24(2): 285–308. <https://doi.org/10.1080/03610919508813243>
74. Beiden SV, Wagner RF, Campbell G (2000) Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. *Acad Radiol* 7(5): 341–349
75. Gallas BD (2006) One-shot estimate of MRC variance: AUC. *Acad Radiol* 13(3): 353–362
76. Hillis SL, Berbaum KS, Metz CE (2008) Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Acad Radiol* 15(5): 647–661
77. Gallas BD, Bandos A, Samuelson FW, Wagner RF (2009) A framework for random-effects ROC analysis: biases with the bootstrap and other variance estimators. *Commun Stat Theory Methods* 38(15):2586–2603. <https://doi.org/10.1080/03610920802610084>
78. Gallas BD, Pennello GA, Myers KJ (2007) Multireader multicase variance analysis for binary data. *J Opt Soc Am A* 24(12):B70–B80
79. Metz CE (1995) The Dorfman/Berbaum/Metz method for testing the statistical significance of ROC differences: validation studies with continuously-distributed data. The Farwest image perception conference to be given October 13, 1995 in Philadelphia, PA
80. Chen W, Wunderlich A, Petrick N, Gallas BD (2014) Multireader multicase reader studies

- with binary agreement data: simulation, analysis, validation, and sizing. *J Med Imaging* (Bellingham) 1(3):031011–031011. <https://doi.org/10.1117/1.JMI.1.3.031011>
81. Obuchowski NA (2009) Reducing the number of reader interpretations in MRMC studies. *Acad Radiol* 16(2):209–217
 82. Obuchowski NA, Gallas BD, Hillis SL (2012) Multi-reader ROC studies with split-plot designs: a comparison of statistical methods. *Acad Radiol* 19(12):1508–1517. <https://doi.org/10.1016/j.acra.2012.09.012>
 83. Chen W, Gong Q, Gallas BD (2018) Paired split-plot designs of multireader multicase studies. *J Med Imaging* (Bellingham) 5(3): 031410. <https://doi.org/10.1117/1.JMI.5.3.031410>
 84. U.S. Food and Drug Administration (2020) Clinical performance assessment: considerations for computer-assisted detection devices applied to radiology images and radiology device data in premarket notification (510(k)) submissions. Guidance for Industry and Food and Drug Administration Staff
 85. Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A (2015) The reusable holdout: preserving validity in adaptive data analysis. *Science* 349(6248):636–638
 86. Gossmann A, Pezeshk A, Wang Y-P, Sahiner B (2021) Test data reuse for the evaluation of continuously evolving classification algorithms using the area under the receiver operating characteristic curve. *SIAM J Math Data Sci* 3: 692–714. <https://doi.org/10.1137/20M1333110>
 87. Hillis SL, Obuchowski NA, Berbaum KS (2011) Power estimation for multireader ROC methods an updated and unified approach. *Acad Radiol* 18(2):129–142
 88. Huang Z, Samuelson F, Tcheuko L, Chen W (2020) Adaptive designs in multi-reader multicase clinical trials of imaging devices. *Stat Methods Med Res* 29(6):1592–1611. <https://doi.org/10.1177/0962280219869370>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 24

Main Existing Datasets for Open Brain Research on Humans

Baptiste Couvy-Duchesne , **Simona Bottani**, **Etienne Camenen**,
Fang Fang, **Mulusew Fikere** , **Juliana Gonzalez-Astudillo** ,
Joshua Harvey , **Ravi Hassanaly**, **Irfahan Kassam** , **Penelope A. Lind** ,
Qianwei Liu, **Yi Lu**, **Marta Nabais**, **Thibault Rolland**, **Julia Sidorenko**,
Lachlan Strike , and **Margie Wright**

Abstract

Recent advances in technology have made possible to quantify fine-grained individual differences at many levels, such as genetic, genomics, organ level, behavior, and clinical. The wealth of data becoming available raises great promises for research on brain disorders as well as normal brain function, to name a few, systematic and agnostic study of disease risk factors (e.g., genetic variants, brain regions), the use of natural experiments (e.g., evaluate the effect of a genetic variant in a human population), and unveiling disease mechanisms across several biological levels (e.g., genetics, cellular gene expression, organ structure and function). However, this data revolution raises many challenges such as data sharing and management, the need for novel analysis methods and software, storage, and computing.

Here, we sought to provide an overview of some of the main existing human datasets, all accessible to researchers. Our list is far from being exhaustive, and our objective is to publicize data sharing initiatives and help researchers find new data sources.

Key words Genetic, Methylation, Gene expression, Brain MRI, PET, EEG/MEG, Omics, Electronic health records, Wearables

1 Aims

We sought to provide an overview and short description of some of the main existing human datasets accessible to researchers. We hope this chapter will help publicize them as well as encourage the sharing of datasets for open science. As much as possible, we tried to provide practical aspects, such as data type, file size, sample demographics, study design, as well as links toward data use/transfer agreements. We hope this can help researchers study larger and more diverse data, in order to advance scientific discovery and improve reproducibility.

This chapter does not aim to provide an exhaustive list of the dataset and data types currently available. In addition, the interested readers may refer to the complementary chapters that focus on data processing, feature extraction, and existing methods for their analyses.

2 Introduction

The availability of data used in research is one of the cornerstones of open science, which contributes to improving the quality, reproducibility, and impact of the findings. In addition, data sharing increases openness and transparent and collaborative scientific practices. The global push for open science is exemplified by the recent publication of UNESCO guidelines [1], the engagement of many research institutions, and the requirements of some scientific journals to make data available upon publication. Finally, the sharing and re-use of data also maximizes the return on investment of the agencies (e.g., states, charities, associations) that fund the data collection.

In light of this, our chapter aims at providing a broad (albeit partial) overview of some of the human datasets publicly available to researchers. To assist researchers and data managers, we first describe the different file formats and the size of the different data types (*see Table 1*). As many of these data are high-dimensional, the size of the data can cause storage and computational challenges, which need to be anticipated before download and analysis. Of note, some datasets cannot be downloaded or analyzed outside of a dedicated system/server. This is the case of the UK Biobank (UKB) exome and whole genome sequencing, whose sheer size has led to the creation of a dedicated Research Analysis Platform, accessible (at some cost) by UKB-approved researchers. In addition, the Swedish registry data is only accessible via national dedicated servers due to the extreme sensitive nature of the data.

This chapter breaks down into sections that focus on each data type, although the same dataset may be mentioned in several sections. Beyond a practical writing advantage (each author or group of authors contributed a section), this also reflects the fact that most datasets are organized around a central data type. For example, the ADNI (Alzheimer's disease Neuroimaging Initiative) focuses on brain imaging and later included genotyping information. Another example is the UKB, which released genotyping data of the 500 K participants in 2017, is now collecting brain MRI (as well as cardiac and abdominal MRI, whole-body DXA, and carotid ultrasound), and has recently made available sequencing data. The different sections also discuss and present the specific data sharing tools and portals (e.g., LONI for brain imaging, GTEx for gene expression) or organization of the different fields (e.g., consortia in

Table 1
Overview of data types and sample size

Data	Subtype	File type/extension	Description	Approx. size of one sample
<i>Clinical</i>				
Self-reports	Text			~100 KB
EHR	Text			~100 KB
<i>Neuroimaging</i>				
MRI	NIfTI (.nii/.nii.gz) or DICOM (.dcm)	3D image (.nii) 2D slices (.dcm)	that compose the 3D volume	T1w ~50 MB T2 FLAIR ~30 MB SWI ~30 MB DWI ~400 MB (highly variable depending on sequence) fMRI ~500 MB (100 MB per minute of acquisition)
PET	NIfTI (.nii / .nii.gz) or DICOM (.dcm)	3D image (.nii) 2D slices (.dcm) that compose the 3D volume		~1 GB after processing
EEG	.edf, .gdf, .eeg, .csv, .mat	Raw EEG signal (.egg, .gdf, .edf – file formats), processed/formatted data (.csv, .mat)		~15 MB
MEG	.fif, .bin, .csv	Raw MEG signal (.fif, .bin) processed/formatted data (.csv, .mat)		~50 MB
<i>Genetics</i>				
Twin/family sample	Text	Phenotypes and pedigree information		~10 KB
Genotyping	.bgen, .bed (.bim + .bed + .fam)	Binary files		~5 MB
GWAS summary statistics	Text	Effect size, test statistic, p-value, effect allele from association testing		Not individual-level data, ~500 MB for a genome-wide association summary
Exome sequencing	.cram, .bcf, .vcf	Each variant site with multiple sequence quality metrics and trained machine learning filters to identify and exclude inconsistencies. Followed by initial QC with various parameters and thresholds		~1 M
Whole genome sequencing	.cram, .bcf, .vcf	Curated using TOPMed variant calling algorithm. All freezes are fully processed		~5 MB

(continued)

Table 1
(continued)

Data	Subtype	File type/extension	Description	Approx. size of one sample
<i>Genomics</i>	Methylation	.idats	Raw binary intensity file (two per sample, one for green and red channels)	~16 MB (450 K)
		.txt or csv.	Processed, normalized, and filtered beta matrix	~27 MB (EPIC)
		.bed (.blim + .bed + .fam)	Fully processed, filtered, and normalized gene expression matrices (in BED format) for each tissue	~8 MB (450 K)
Expression		Text	Covariates used in eQTL analysis. Includes genotyping principal components and PEER factors	~14 MB (EPIC) ~1 MB per individual ~6 KB
<i>Smartphone and sensors</i>	<i>Actigraphy</i>	Varies with product (e.g., GENEActiv uses .bin)	Raw accelerator data	~500 MB (GENEActiv, 14 days of recording)

Size of one sample may vary based on the technology used to generate the data (e.g., MRI resolution, genotyping chip)

genetics). Every time, we have tried to include the largest dataset(s) available, as well as the commonly used ones, although the selection may be subjective and reflect the authors' specific interests (e.g., age or disease groups).

All datasets are listed in a single table (*see* Table 2), which includes information about country of origin, design (e.g., cross-sectional, longitudinal, clinical, or population sample), and age range of the participants. Unless specified, the datasets presented include male and female participants, although the proportion may differ depending on the recruitment strategy and disease of interest. In addition, the table lists (and details) the different data types that have been collected on the participants. We have only focused on a handful of data types: genetic data (including twin/family samples, genotyping, and exome and whole genome sequencing), genomics (methylation and gene expression), brain imaging (MRI and PET), EEG/MEG, electronic health records (hospital data and national registry), as well as wearable and sensor data. However, we have included additional columns "Other omics" and "Specificities" that list other types of data being collected, such as proteomics, metabolomics, microRNA, single-cell sequencing, microbiome, and non-brain imaging.

Our main table (*see* Table 2) also includes the URLs to the dataset websites and data transfer/agreement. From our experience, data access can take between an hour and up to a few months. The agreements almost always require a review of the project and to acknowledge the data collection team and funding sources (e.g., under the form of a byline, a paragraph in the acknowledgment, and more rarely co-authorships). Standard restrictions of use include that the data cannot be redistributed and that the users do not attempt to identify participants. Specific clauses are often added depending on the nature of data and the specific laws and regulations of the countries it originates from.

There is a growing scientific and ethical discussion about the representativity of the datasets being used in research. Researchers should be aware of the biases present in some datasets (e.g., "healthy bias" in the UKB [2]), which should be taken into account in study design (e.g., analysis of diverse ancestry being collected in genetics [3]), when reporting results [2, 4] and evaluating algorithms [5, 6]. Overall, our (selected) list exemplifies the need for datasets from under-represented countries or groups of individuals (e.g., disease, age, ancestry, socioeconomic status) [7, 8]. Our main table (*see* Table 2) will be accessible online, in a user-friendly, searchable version. Finally, we will also make this table collaborative (via GitHub <https://github.com/baptisteCD/MainExistingDatasets>) in order to grow this resource beyond this book chapter.

Table 2
Description of a selection of the main human datasets available for research

Sample	N	Age range	Country	Cross-sectional/ longitudinal	Clinical	Neuroimaging (MRI, PET, fMRI)	EEG/MEG	Genetics (genotyping, exome, WGS, twins)	Genomics	Smartphones and sensors	Other omics	Website/Data Transfer Agreement	Reference article(s)	Specificities
<i>Alzheimer's Disease Neuroimaging Initiative (ADNI)</i>	2215	55-90	USA and Canada	Longitudinal (up to 9 years of follow-up, ongoing)	MRI (T1w, T2, DWI, rsfMRI) PET (¹⁸ F-FDG, FBB, AV45, PIB)	NA	Genotyping, WGS	Methylation (subset of 653 individuals— 3 time points)	Transcriptomics, CSF proteomics, metabolomics, lipidomics	Four waves (ADNI), 2, GO, 3, with different inclusion and protocols	http://adni.loni.usc.edu/ http://data-samples.access-data.jp	[11, 19]		
<i>Australian Imaging, Biomarkers and Lifestyle Study of Ageing (AIBL)</i>	726	60+	Australia	Longitudinal (up to 6 years of follow-up)	MRI (T1w, T2, DWI, rsfMRI) PET (PIB, AV45, Flutre)	NA	Genotyping	Methylation	ActiGraph activity (10% of sample)	https://ida.loni.usc.edu/ https://collaboration.access-applicense.jp	[12]	Neuroimaging and selected clinical data available via the LONI platform Full sample (N= 1200), including extended clinical, genotyping, or methylation available via application to CSIRO		
<i>Open Access Series of Imaging Studies n3 (OASIS3)</i>	1096	42-95	USA	Longitudinal (up to 12 years of follow-up)	MRI (T1w, T2w, FLAIR, ASL, SWI, time of flight, rsfMRI, and DWI)	NA	NA	NA	NA	https://www.oasis-brain.org https://www.oasis-brain.org/#access	[13]	Retrospective dataset from imaging projects collected by WUSTL Knight ADRC over 30 years Two other (non-independent) datasets available: OASIS1 and OASIS2		
<i>Adolescent Brain Cognitive Development (ABC-D)</i>	-11,878	9-12	USA	Longitudinal (up to 2 years of follow-up, ongoing)	MRI (T1w, T2, rsfMRI, tfMRI)	NA	Genotyping, pedigree (twinning)	iPad tasks and testing	NA	https://abcdstudy.org https://nda.nih.gov/abcd/request-access	[16, 17]	Objective of 10 years of follow-up		
<i>Enhancing Neuroimaging Genetics through Meta- analysis (ENIGMA)</i>	>50,000 indiv. incl.: 9572 (SCZ), 6503 (BD), 10,105 (MDD), 18,688 (PTSD), 3240 (SUD), 3665, (OCD), 41,801 (ADHD), 18,605 (life span)	3-90 (no restriction)	Worldwide (43+ countries)	Cross-sectional and longitudinal	Psychiatry, neurology, addiction, suicidality, brain injury, HIV, antisocial behavior	Resting-state EEG	Genotyping, CNVs, pedigree (twinning)	Methylation	NA	https://enigma.mindtrc.org http://join.enigma-initiatives.org http://research.enigma-platform.org/download-enigma-gwas-results/	[27] [112, 114]	Neuroimaging Consortium organized around genetics and/or disease/ trait working groups as well as non-clinical working groups with focus on sex, healthy aging, plasticity, etc. Imaging and generic protocols and genome-wide association statistics are available for download on the ENIGMA website		

<i>ChiNirr Brain PET 116 Template (CNPET)</i>	27-81	China	Cross-sectional	Healthy subjects	PET (¹⁸ F-FDG)	NA	NA	NA	NA	https://www.nitrc.org/doi/landing_page. PHP?table_group&id=148&&doi=https://www.nitrc.org/projects/cnpet	[39]	
<i>Centre for Integrated Molecular Brain Imaging (CIMBI)</i>	-2000	17-93	Denmark	Cross-sectional	Mental and physical state, personality, and background	MRI (T1w, T2w, DWI (MRI) PET (¹¹ C-5-HT))	NA	Genetic polymorphisms relevant for the 5-HT system	NA	NA	Data sharing only within the European Union (EU). Visiting researchers for non-EU	
<i>Motor Imagery dataset from Cho et al. [192]</i>	52	24.8 SD = 3.86 years old	South Korea	Cross-sectional	Healthy subjects	NA	EEG (64 channels); real hand movements, motor imagination hand movement tasks	NA	NA	http://www.cimbi.dk/index.php	[38]	
<i>EEG Alpha Waves dataset</i>	20	19-44 years old	France	Cross-sectional	Healthy subjects	NA	EEG (16 channels); resting state, eyes open/ closed	NA	NA	http://mabb.netrotch.com/docs/generated/mabb_datasets_Cho2017.html#mabb_datasets.Cho2017	[192]	
<i>Multisensor spectra recorded during GA/Anesthetic unconsciousness</i>	55	USA	Cross-sectional	Healthy participants and patients receiving an anaesthesia care in an operating room context	NA	EEG (64 for healthy volunteers, 6 for patients under anaesthesia)	NA	NA	NA	https://zenodo.org/record/2605110#.YTEt5Z4zZhE	[193]	
<i>A multi-subject, multi-modal human neuroimaging dataset</i>	19 (16 for MEG and EEG, 3 for fMRI)	23-37 years old	UK	Longitudinal (2 visits over 3 months)	Healthy subjects	fMRI	MEG, EEG (70 channels); resting state, sensory-motor tasks	NA	NA	https://physionet.org/content/cog-power-anesthesia/1.0.0/	[194]	
<i>Motor imagery induced classifier application</i>	10	26-46 years old	Germany	Cross-sectional	Healthy subjects	NA	EEG (59 channels); hands, feet, and tongue motor imagination tasks	NA	NA	https://legacy.openfmri.org/dataset/ds000117/	[195]	
<i>BCI Competition 2008-2009 dataset A</i>	9	NA	Austria	Cross-sectional	Healthy subjects	NA	EEG (22 channels); hands, feet, and tongue motor imagination tasks	NA	NA	http://www.bbci.de/competition/w/desc_1.html	[196]	
										Data used for a BCI competition. Also include simulated/synthetic data		
										Data used for a BCI competition	[197]	

(continued)

Table 2
(continued)

Sample	N	Age range	Country	Cross-sectional/longitudinal	Clinical	Neuroimaging (MRI, PET, fMRI)	EEG/MEG	Genetics (genotyping, exome, WGS, twins)	Smartphones and sensors	Other omics	Website/Data Transfer Agreement	Reference article(s)	Specificities
<i>BCI Competition 2008–Graz data set B</i>	9	NA	Austria	Cross-sectional	Healthy subjects	NA	EEG (3 channels)	NA	NA	NA	http://www.bbci.de/competition/viv-desc22b.pdf	[197]	Data for a BCI competition
<i>EEG dataset from Munitz et al.</i> [198]	64 (34 MDD, 30 healthy) years old	Malaysia	Longitudinal (multiple visits to the clinic)	Case control form major depressive disorder	NA	EEG (19 channels); hands/motor imagination tasks	NA	NA	NA	NA	https://journals.plos.org/biosone/article?id=10.1371/journal.pone.0171409	[198]	EEG (resting task), MDD based on medical history of patients
<i>EEG data for ADHD control children</i>	121	7–12 years old	South Korea	Cross-sectional	ADHD (61 children) and healthy controls [61]	EEG (19 channels); visual attention tasks	NA	NA	NA	NA	https://icee-dataport.org/open-access/eeg-data-adhd-control-children	[199]	Visual cognitive tests on children with ADHD, using videos
<i>EEG and EOG data from Grimaldi-Gonzalez et al.</i> [61]	NA	Germany	Longitudinal (2 to 10 visits)	Amyotrophic lateral sclerosis, locked-in state	EEG (116 channels)	EEG (116 channels)	NA	NA	NA	NA	https://doi.org/10.6084/m9.figshare.13148762	[200]	Spelling task with eye movement-based answers
<i>Temple University Hospital (TUH) EEG Corpus</i>	10,874	1–90	USA	Cross-sectional	Epilepsy; stroke, concussion, healthy	EEG (20–41 channels)	NA	NA	NA	NA	https://sip.picopenpress.com/projects/tuh_eeg/html/downloads.shtml	[201]	Extensive dataset of 30 K clinical EEG recordings collected at TUH from 2002 to today. Epilepsy subset: TUEP dataset
<i>Bern-Barcelona EEG database</i>	5	NA	Spain	Longitudinal	Epilepsy	NA	EEG (64 channels); rest	NA	NA	NA	https://repository.tuh.ch/handle/10330/42829	[202]	Intracranial EEG samples before and after surgery
<i>CHB-MIT Scalp EEG Database</i>	22	1.5–22	USA	Longitudinal	Seizures and intractable seizures	NA	EEG (21 channels); resting state	NA	NA	NA	https://www.physionet.org/content/chbmit/1.0.0/	[203]	Pediatric subjects with intractable seizures
<i>Brain/Neural-Computer Interaction (BNCI) Horizon</i>	2	NA	Austria	Longitudinal	Chronic stroke	NA	EEG (2 channels); eye staring task	NA	NA	NA	http://bnchi-horizon.2020.eu/database-data-sets	[204]	The BNCI is an open-source project with several datasets. Stroke is “e. SCP training in stroke (06-2014).”
<i>Queensland Twin Adolescent Brain (QTAB)</i>	422 (baseline)	9–14 (baseline)	Australia	Longitudinal	Population sample	MRI (T1w, T2w, FLAIR, DWI, rsfMRI, tMRI, ASL)	Pedigree (twin/sibling), genotyping	NA	Wrist-worn accelerometer	https://imgenomics.netau/	[205]	Includes participants from Queensland Twin Registry and Twins Research Australia	

<i>Queensland Twin Imaging Study (QTS)</i>	1200+	12–30	Australia	Cross-sectional	Population sample (as part of BATS). Self-report mental health, cognition, substance use, and personality measures	MRI (T1w, DWI, rsfMRI, tMRI)	Pedigree (twin/ sibling), genotyping	NA	NA	NA	https://imaginggenomics.net.au/	[206]	Include participants from Brisbane adolescent twin study (BATS)
<i>Human Connectome Project, young adults (HCP-YA)</i>	-1200	22–35	USA	Cross-sectional	Population sample. Self-report mental health, cognition, personality, and substance use measures	MRI (T1w, T2w, DWI, rsfMRI, tMRI)	Pedigree (twin/ sibling). Genotyping	NA	NA	NA	https://db.humanconnectome.org	[29]	Expanded to development and aging projects (similar imaging protocols, but extensions do not include twins)
<i>Vietnam Era Twin Study of Aging (VETSA)</i>	12–37	51–60 (baseline; 500+ with MRI)	USA	Longitudinal	US veterans, males only	MRI (T1w, DWI, ASL (subset of sample))	Pedigree (twin). Genotyping	NA	NA	NA	https://mcschool.psychiatry.som.yale.edu/som/~psychiatry/research/VETSA/Researchers/Pages/default.aspx	[83]	Subset of the Vietnam Era Twin Registry, males only
<i>Older Australian Twins Study (OATS)</i>	62.3 (baseline)	>65	Australia	Longitudinal	Population sample. Self-report medical and mental health and neuropychological measures	MRI (T1w, DWI, tMRI); PET	Pedigree (twin/ sibling). Genotyping	NA	NA	NA	https://cheba.unsw.edu.au/research-projects/older-australian-twins-study	[84]	Recruited through the Australian Twin Registry
<i>Swedish Twin Registry (STR)</i>	87,000 twin pairs	All ages	Sweden	Longitudinal (since the late 1950s)	NA	NA	Genome-wide single nucleotide polymorphism array	Methylation is not available in the STR yet but is available in the Swedish genotyping array	NA	NA	https://kisec/en/research/the-swedish-twin-registry	[74]	Twin Registry
<i>UK Biobank (UKB)</i>	-502,000 -50,000 (imaging subset) -100,000 (actigraphy subset)	37–73 44–82 (imaging) 40–69 (actigraphy)	UK	Longitudinal	Self-reported and EHR medical history (incl. cancer, neurology, COVID-19)	MRI (T1w, T2w, FLAIR, DWI, SWI, rsfMRI, tMRI)	Genotyping, exome, WGS, pedigree	Actigraphy	Metabolomics (future release)	https://biobank.ndph.ox.ac.uk/ams/rexApplications	[14, 101, 186]	Population based sample (volunteers), healthy bias, ongoing resource collection/target MRI sample 100 K, rest 10 K, also available: Cardiac and abdominal MRI, whole-body DXA, dual-energy X-ray absorptiometry,	

(continued)

Table 2
(continued)

Sample	N	Age range	Country	Cross-sectional/ longitudinal	Clinical	Neuroimaging (MRI, PET, fMRI)	EEG/MEG	Genetics (genotyping, exome, WGS, twins)	Genomics	Smartphones and sensors	Other omics	Website/Data Transfer Agreement	Reference article(s)	Specificities
														carotid ultrasound, electrocardiogram, hematology assays, serological antibody response assay, telomere length
<i>Avon Longitudinal Study of Parents and Children (ALSPAC): Accessible Resource for Integrated Epidemiologic Studies (ARIES)</i>	2044	Average age: Mothers (antenatal = 28.7, follow-up = 47.5), offspring (birth = 40 weeks, childhood = 7.5, adolescence = 17.1)	UK	Longitudinal (2 time points for mother, 3 for offspring)	Clinical evaluations, obstetric data, cognition questionnaire	MRI (T1w, DWI, mcDESPOT (subset of offspring cohort))	Genotyping	Methylation	Transcriptomics	http://www.aresepigenomics.org.uk/ https://github.com/MRCIEU/aries	[207]	General population study (health and development) following 1022 mother-offspring pairs; 2 time points for mother, 3 for offspring		
<i>Biobank-based Integrative Omics Studies (The BIOS Consortium)</i>	~4000	18-87	Netherlands	Longitudinal	Clinical information and biosays (depending on the sub-cohort)	NA	Genotyping, pedigree	Methylation	Transcriptomics, metabolomics	https://www.bbmm.nl/ node/24	[208]	Population study including various sub-cohorts, covering differing research designs (LifeLines, Leiden longevity study, Netherlands twin registry, Rotterdam study, CODAM, and the prospective ALS study Netherlands); access via European genome-phenome archive (EGA) and SURSeia high performance computing cloud		
<i>Framingham Heart Study</i>	4241	Offspring cohort	USA	Cross-sectional (multi-generational)	Extensive clinical evaluations and biosays, original focus on cardiovascular diseases	MRI (T2w)	NA	Genotyping, WGS	Methylation	Transcriptomics, metabolomics	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs0000724.v9.p13	[209, 210]	Data collected over two generations of individuals; cardiovascular MRI A subset of individuals are used to identify rare variants influencing brain imaging phenotypes. Data included in NHIBI TOPMed	

<i>Women's Health Initiative</i>	2129	50–79	USA	Cross-sectional	Clinical evaluations and bioassays; cardiovascular diseases, women only	NA	Genotyping	Methylation	NA	Transcriptomics, metabolomics, mRNA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001335.v2.p3	[211]	Women only
<i>Sporadic ALS Australia Systems Genetics Consortium (SALS-AGC)</i>	1395	Predicted mean age from DNA methylation	Australia	Cross-sectional	Case control of amyotrophic lateral sclerosis	NA	Genotyping	Methylation	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002068.v1.p1	[212]	
<i>System Genomics of Parkinson's Disease (SGPD)</i>	2333	Predicted mean age from DNA methylation	Australia and New Zealand	Cross-sectional	Case control of Parkinson's disease	NA	Genotyping	Methylation	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?acc=GSE145361	[213]	
<i>Genetics of DNA Methylation Consortium (gDMC)</i>	32,851	0–91	Worldwide	Cross-sectional	Multiple diseases, but also healthy aging cohorts	NA	Genotyping	Methylation	NA	NA	http://www.godmc.org.uk/cohorts.html	[79]	Consortium gathering 38 independent studies; data access to be obtained from each subsample
<i>Psychiatric Genomics Consortium (PGC)</i>	By 2025, ~2.5 million cases of psychiatric disorders	All ages	Worldwide	Cross-sectional	Psychiatric disorders, substance use disorders, and neurology: Major depressive disorder, cannabis use disorder, alcohol use disorder, schizophrenia, anorexia, bipolar, ADHD, Alzheimer's	NA	Genotyping and sequencing	Expression and methylation data available in some working groups	NA	NA	https://www.mged.info/ctd/	[214, 215]	Consortium organized around disease/ trait working groups
<i>The genetic epidemiology of asthma in Costa Rica</i>	4347	6–12	Costa Rica	Cross-sectional	Asthma cases and controls	NA	WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA295246	[216]	A subset of individuals from this sample/ pharmacogenetic drug response in children with asthma. Data included in NHLBI T0PMed
<i>Coronary Artery Risk Development in Young Adults (CARDIA)</i>	3425	18–30	USA	Longitudinal	Coronary Artery Risk	NA	WGS	NA	NA	NA	https://anvilproject.org/neuro/data/studies/phs001612	[217]	A subset of longitudinal dataset of 3087 self-identified Black and White participants from the CARDIA study were used to study multi-ethnic polygenic risk score

(continued)

Table 2
(continued)

Sample	N	Age range	Country	Cross-sectional/ longitudinal	Clinical	Neuroimaging (MRI, PET MRI)	EEG/MEG	Genetics (genotyping, exome, WGS, twins)	Genomics	Smartphones and sensors	Other omics	Website/Data Transfer Agreement	Reference article(s)	Specificities
<i>Genetic Epidemiology Network of Arteriopathy (GENOA)</i>	1854	>60	USA	Longitudinal	Elucidate the genetics of target organ complications of hypertension	NA	WGS	NA	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001345.v3.p1	[217]	Study was used to study multi-ethnic polygenic risk score associated with hypertension prevalence and progression
<i>Hispanic Community Health Study/Study of Latinos (HCHS/SOL)</i>	8093	18-74	USA	Longitudinal	A multicenter prospective cohort study for asthma patients	NA	WGS	NA	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001395.v1.p1	[217]	Study was used to study multi-ethnic polygenic risk score associated with hypertension prevalence and progression
<i>Women's Health Initiative (WHI)</i>	11,357	>65	USA	Longitudinal	Women's Health Initiative cohort involved study on ischaemic stroke, 900 cases of hemorrhagic stroke	NA	WGS	NA	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs00200.v12.p3	[217]	Study was used to study multi-ethnic polygenic risk score associated with hypertension prevalence and progression
<i>Atherosclerosis Risk in Communities (ARIC)</i>	8975	45-64	USA	Cross-sectional/ longitudinal	Red blood cell phenotype	NA	WGS	NA	NA	NA	NA	https://anvilproject.org/datasets/phs001211	[218]	WGS association study of red blood cell phenotypes, GWAS statistics available. Data included in NHLBI TOPMed
<i>Rare Variants for Hypertension in Taiwan Chinese (THRV)</i>	2159	>35	Taiwan/China, Japan	Longitudinal	Insulin-resistant cases and controls	NA	WGS	NA	NA	NA	NA	https://anvilproject.org/neu/data/studies/phs001387	[219]	Clustering and heritability of insulin resistance in Chinese and Japanese hypertensive families. Data included in NHLBI TOPMed
<i>My Life, Our Future initiative (MLOF)</i>	7482	>18	USA	Cross-sectional	Hemophilia cases and controls	NA	WGS	NA	NA	NA	NA	https://atm.org/what-we-do/national-projects/molof-research-repository.html	[220]	Summary statistics, different types of DNA variants detected in hemophilia. Data included in NHLBI TOPMed

<i>Genetic Epidemiology of COPD (COPDGene)</i>	19,996	45-80	USA	Cross-sectional	Pulmonary functions	NA	NA	WGS	NA	NA	Gene expression (eQTL) and methylation (mQTL); eQTLs in 48 tissues from GTEx v7	https://anvilproject.org/data/studies/ph001211	[221]	Multi-omic data from GTEx and TOPMed identify potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
<i>Cardiovascular Health Study (CHS)</i>	4877	>65	USA	Longitudinal	<i>Cardiovascular health</i>	NA	NA	WGS	NA	NA	NA	https://ega-archive.org/studies/ph001368/ DOI:https://doi.org/10.1038/e41467-020-18334-7	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
<i>Cleveland Family Study (CFS)</i>	3576	>65	USA	Longitudinal	Epidemiological data on genetic and non-genetic risk factors for sleep disordered breathing	NA	NA	WGS	NA	NA	NA	https://ega-archive.org/studies/ph000954/ DOI:https://doi.org/10.1038/e41467-020-18334-7	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
<i>Framingham Heart Study (FHS)</i>	4241	<65	USA	Longitudinal	Assess risk of cardiovascular disease study	NA	NA	WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=ph000724+9.p13	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
<i>Jackson Heart Study (JHS)</i>	3596	>55	USA	Longitudinal	Assess cardiovascular disease	NA	NA	WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=ph000964+5.p1	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
<i>Multi-Ethnic Study of Atherosclerosis (MESA)</i>	6814	45-84	USA	Longitudinal	Assess cardiovascular disease	NA	NA	WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/dataset/dbgap/phys001416	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
<i>Boston Early-Onset COPD (ECOARD)</i>	80	<53	USA	Longitudinal	Chronic obstructive pulmonary disease (COPD)	NA	NA	WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=ph000946.v5.p1	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
<i>Genetic Epidemiology of COPD (COPDGene)</i>	10,647	45-80	USA	Longitudinal	Chronic obstructive pulmonary disease (COPD)	NA	NA	WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=ph000951.v5.p5	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed

(continued)

Table 2
(continued)

Sample	N	Age range	Country	Cross-sectional/longitudinal	Clinical	Neuroimaging (MRI, fMRI)	EEG/MEG	Genetics (genotyping, exome, WGS, twins)	Smartphones and sensors	Other omics	Website/Data Transfer Agreement	Reference article(s)	Specificities
<i>Genotype-Tissue Expression (GTEx)</i>	948 individuals, 20–70 17,382 samples (across 54 tissues)	USA	Cross-sectional	General medical history, HIV status, potential exposures, medical history at the time of death, death circumstances, serology results, smoking status	NA	WGS	Gene expression (bulk RNA-seq)	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/dataset.cgi?study_id=phs000424.v8_p2&phv=169091&phd=39108&phf=8&phdf=&phl=&phm=&phn=&phf=&phm=&phl=&consent=8&temp=-1	[222]	Postmortem samples
<i>Developmental Genotype-Tissue Expression (dGTEs)</i>	120 individuals, 0–18 3,600 samples (across 30 tissues)	USA	Cross-sectional	General medical history	NA	WGS	Gene expression (bulk RNA-seq)	NA	NA	NA	https://digrex.org/	NA	Expected data release by late 2022
<i>Greater-Paris Area Hospital (AP-HP)</i>	Several millions All ages ~130 K patients with brain MRI (~200 K images)	France	Longitudinal	ICD10 codes for diagnoses, medical reports, and a wide range of physiological variables	All medical imaging modalities and all organs: MRI, PET, SPECT, CT, ultrasounds	NA	NA	NA	NA	NA	https://cds.ap-hp.fr/	[155]	All patients visiting greater-Paris area hospital (AP-HP) the figures correspond to a specific project approved on the clinical data warehouse (CDW)
<i>Swedish Total Population Register (TPR)</i>	Nationwide coverage (N = 9,775,572 on April 30, 2015)	All ages	Sweden	Longitudinal (since 1968)	NA	NA	NA	NA	NA	NA	www.scb.se/varantester/bestaldan-och-statistik/besfalls-mikrodata-vilka-mikrodata-firms-individregister/regster-over-tidslinjer-kningens-rtb/	[158]	Demographic information

<i>Swedish Longitudinal Integrated Database for Health, Insurance and Labour Market Studies (LISA)</i>	All individuals ≥16 years in Sweden	≥16 years since 2010	Sweden	Longitudinal (since 1990)	NA	NA	NA	NA	NA	NA	NA	www.scb.se/lisa	[159]	Demographic and socioeconomic information
<i>Swedish Multi-Generation Register (MGR)</i>	Over 11 million All ages	Sweden	Longitudinal (since 1961)	NA	NA	NA	Pedigree	NA	NA	NA	NA	www.scb.se/en/-finding-statistics/statistics-by-subject-area/other-other-publications/non-statistical-publications/multi-generations-register-2016/	[160]	Family relation
<i>Swedish National Patient Register (NPR)</i>	Nationwide coverage	All ages	Sweden	Longitudinal (since 1964)	Nationwide inpatient care since 1987 and outpatient care since 2001	NA	NA	NA	NA	NA	NA	www.socialstyrelsen.se/en/statistics-and-data/registers/register-information/the-national-patient-register/	[161]	Clinical diagnoses from inpatient and outpatient care
<i>Swedish Cancer Register (SCR)</i>	Around 60,000 All ages malignant cases are included annually (statistics in 2020)	Sweden	Longitudinal (since 1958)	Clinical diagnoses of cancer	NA	NA	NA	NA	NA	NA	NA	www.socialstyrelsen.se/en/statistics-and-data/registers/register-information/swedish-cancer-register/	[165]	
<i>Swedish Medical Birth Register (MBR)</i>	Infants at birth 80,000-120,000 deliveries annually	Sweden	Longitudinal (since 1973)	Yes	NA	NA	NA	NA	NA	NA	NA	https://www.socialstyrelsen.se/en/statistics-and-data/registers/register-information/the-swedish-medical-birth-register/	[167]	Register of medical birth
<i>Swedish Causes of Death Register (CDR)</i>	Neatly 100,000 All ages deaths annually	Sweden	Longitudinal (since 1952)	Cause of death	NA	NA	NA	NA	NA	NA	NA	https://www.socialstyrelsen.se/statistik-odh-data/register/all-doktors-kregister/	[170]	Register of cause of death
<i>Swedish Prescribed Drug Register (PDR)</i>	More than 100 million records annually	Sweden	Longitudinal (since July 2005)	Prescribed drug(s)	NA	NA	NA	NA	NA	NA	NA	https://www.socialstyrelsen.se/en/statistics-and-data/registers/register-information/the-swedish-prescribed-drug-register/	[172]	Register of prescribed drug

(continued)

Table 2
(continued)

Sample	N	Age range	Country	Cross-sectional/ longitudinal	Clinical	Neuroimaging (MRI, PET, fMRI)	EEG/MEG	Genetics (genotyping, exome, WGS, twins)	Genomics	Smartphones and sensors	Other omics	Website/Data Transfer Agreement	Reference article(s)	Specificities
<i>Swedish Dementia Registry (SDR)</i>	More than 100,000	27–103 (between 2007 and 2012)	Sweden	Longitudinal (since 2007)	Dementia	MRI, PET, CT	EEG	NA	NA	NA	NA	www.svdemnet.se	[176]	Register of dementia
<i>Swedish Neuro-Register (SNR)</i>	Around 16,000 All ages	Sweden	Longitudinal (since 2004)	Motor neuron disease, MS	MRI, PET, and CT	MEG	Genotyping	NA	NA	NA	NA	www.natorg.se	[177]	Register of MND (especially MS)
<i>Swedish Stroke Register (Riks-Stroke)</i>	Around 29,000 Mean age 75 for Sweden cases (around 21,000 stroke and 8000 TIA) annually (statistics in 2020)	Sweden	Longitudinal (since 1994)	Stroke and transient ischemic attack (TIA)	MRI and CT scan	NA	NA	NA	NA	NA	NA	www.risksstroke.org/	[179]	Register of stroke and TIA; includes electrocardiograms
<i>Brisbane Adolescent Twin Sample (BATS)</i>	>11	Australia	Cross-sectional and longitudinal	Population sample. Self-report mental health, cognition, substance use, and personality measures	EEG (15 channels, eyes closed resting; N ~ 1000)	Pedigree (twin/sibling), genotyping	NA	Wrist-worn accelerometer (N = 130)	NA	NA	https://imaginggenomics.net/	[79, 223–225]	Also known as the Brisbane Longitudinal Twin Study (BLTS). Includes participants from the Queensland Twin Registry	
<i>mPower</i>	~8000	>18	USA	Longitudinal	Parkinson's disease (subsample self-identified professional diagnosis)	NA	NA	NA	NA	iPhone application	NA	https://parkinsonpower.org/team https://www.synapse.org/#Synapse499293/wik/247860	[226]	Sample size varies across surveys and tasks completed

We hope this overview could be useful to the readers wanting to replicate findings, maximize sample size and statistical power, develop and apply methods that utilize multi-level data, or even select the most relevant dataset to tackle a research question. We also hope this encourages the collection of new data shared with the community while ensuring interoperability with the existing datasets.

3 Neuroimaging

3.1 Magnetic Resonance Imaging (MRI)

Brain magnetic resonance images are 3D images that measure brain structure (T1w, T2w, FLAIR, DWI, SWI) or function (fMRI). The different MRI sequences (or modalities) can characterize different aspects of the brain. For example, T1w and T2w offer the maximal contrast between tissue types (white matter, gray matter, and cerebrospinal fluid), which can yield structural/shape/volume measurements. They can also be used in conjunction with an injection of a contrast agent (e.g., gadolinium) for detecting and characterizing various types of lesions. FLAIR is also useful for detecting a wide range of lesions (e.g., multiple sclerosis, leukoaraiosis, etc.). SWI focuses on the neurovascular system, while DWI allows measuring the integrity of the white matter tracts. Functional MRI measures BOLD (blood oxygen level dependent) signal, which is thought to measure dynamic oxygen consumption in the different brain regions. Of note, fMRI consists of a series of 3D images acquired over time (typically 5–10 min).

Brain MRI is available as a series of DICOM files (brain slices, traditional format of the MRI machines) or a single NIfTI (single 3D image) format (*see Table 1*). The two formats are roughly equivalent, and most image processing pipelines allow both data sources as input. MR images are composed of voxels (3D pixel), and their size (e.g., 1 × 1 × 1 mm) corresponds to the image resolution.

In practice, most MR images are archived and shared via web-based applications and more rarely using specific software (e.g., UKB). The two major web platforms are XNAT (eXtensible Neuroimaging Archive Toolkit) [9], an open-source platform developed by the Neuroinformatics Research Group of the Washington University School of Medicine (Missouri, (1, 2)), and IDA (Image and Data Archive) created by the Laboratory of NeuroImaging of the University of South California (LONI, <https://loni.usc.edu/>). Of note, XNAT also allows to perform some image processing [9].

The neuroimaging community has developed BIDS (Brain Imaging Data Structure), a standard for MR image organization to accommodate multimodal acquisitions and facilitate processing.

In practice, few datasets come in BIDS format, and tools have been developed to assist with download and conversion (e.g., <https://clinica.run>) [10].

We have listed a handful of datasets (*see Table 1*), which is far from being exhaustive but aims at summarizing some of the largest and/or most used samples. Our selection aims at presenting diverse and complementary samples in terms of age range, populations, and country of origin.

First, we have described three clinical elderly samples from the USA and Australia, with a focus on Alzheimer's disease and cognitive disorders. The Alzheimer's Disease Neuroimaging Initiative (ADNI) was launched in 2004 and funded by a partnership between private companies, foundations, the National Institute of Health, and the National Institute for Aging. ADNI is a longitudinal study, with data collected across 63 sites in the USA and Canada. To date, four phases of the study have been funded, which makes ADNI one of the largest clinical neuroimaging samples to study Alzheimer's disease and cognitive impairment in aging. ADNI collected a wide range of clinical, neuropsychological, cognitive scales as well as biomarkers, in addition to multimodal imaging and genotyping data [11]. Sites contribute data to the LONI, which is automatically shared with approved researchers without embargo. The breadth of data available and its accessibility have made ADNI one of the most used neuroimaging samples, with more than 1000 scientific articles published to date.

The Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) started in 2006 and has since recruited about 1100 participants over 60 years of age, who have been followed over several years (*see Table 1*) [12]. AIBL collected data across the different Australian states and, similar to ADNI, consisted in an in-depth assessment of individual cognition, clinical status, genetics, genomics, as well as multimodal brain imaging [12]. In 2010, AIBL partnered with ADNI to release the AIBL imaging subset and selected clinical data via the LONI platform. Having the same MRI protocols and similar fields collected, AIBL represents a great addition to the ADNI study, by boosting statistical power or allowing for replication. The full clinical information as well as genetics, genomics, and wearable data (actigraphy watches) are not available via the LONI and require a direct application to the Commonwealth Scientific and Industrial Research Organisation (CSIRO) (*see Table 1*).

The Open Access Series of Imaging Studies v3 (OASIS3) is another longitudinal sample comprising almost 1100 adult participants (*see Table 1*) [13]. Its main focus is around aging and neurological disorders, and the application/approval process is extremely fast (typically a couple of days). OASIS3 is hosted on XNAT and is the third dataset to be made available by the Washington University in Saint Louis (WUSTL) Knight Alzheimer's Disease Research

Center (ADRC), although the three datasets are not independent and cannot be analyzed together. Contrary to ADNI and AIBL, OASIS3 is a retrospective study that aggregates several research studies conducted by the WUSTL over the past 30 years. As a result, the data collected may vary from one individual to the next, with a variable time window between visits. In that sense, OASIS3 resembles data from clinical practice, with individual specific care/assessment pathways.

The UK Biobank (UKB) imaging study [14] is the largest brain imaging study to date, with around 50,000 individuals already imaged (target of 100,000). The imaging wave complements the wealth of data already collected in the previous waves (*see* Table 1; *see also* Subheading 5 for a description of the full dataset). Considering the sheer size of the data, the biobank shares raw and processed images as well as structured data (measurements of regions of interest) [15]. Data is accessible upon request by all bona fide researchers, with certified profiles. Data access requires payment of a fee, which only aims to cover the biobank functioning costs. The UKB has developed proprietary tools for a secure download and data management (<https://biobank.ndph.ox.ac.uk/showcase/download.cgi>).

The Adolescent Brain Cognitive Development (ABCD) is an ongoing longitudinal study of younger individuals, recruited aged 9–10 years and who will be followed over a decade [16, 17]. The ABCD focuses on cognition, behavior, and physical and mental health (e.g., substance use, autism, ADHD) of adolescents. It includes self- and parental rating of the adolescents as well as a description of the familial environment [17]. ABCD data is hosted on the NIMH data archive and requires obtaining and maintaining an NDA Data Use Certification, which requires action from a signing official (SO) from the researcher institution, as defined in the NIH eRA Commons (<https://era.nih.gov/files/eRA-Commons-Roles-10-2019.pdf>).

The Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) disease working groups have stemmed from the ENIGMA genetics project (*see* Subheading 5.3) to perform worldwide neuroimaging studies for a wide range of disorders (e.g., major depressive disorder [18], attention-deficit hyperactivity disorder [19], autism [20], post-traumatic stress disorder [21], obsessive-compulsive disorder [22], substance dependence [23], schizophrenia [24], bipolar disorder [25]) as well as traits of interest (e.g., sex, healthy variation [26]); *see* [27] for a review. Each working group may conduct simultaneously several research projects, proposed and led by its members. Each site of the working group chooses the project(s) they contribute to and performs the analyses. Of note, most ENIGMA working groups still rely on a meta-analytic framework, even if recent projects (e.g., machine learning) now require sharing data onto a central server. Interested

researchers can contribute new data and propose analyses or new image processing pipelines to the different working groups. The ENIGMA samples typically comprise thousands of participants (controls and/or cases; *see* Table 1), and data are inherently heterogeneous, each site having specific recruitment and protocols.

Other neuroimaging MRI datasets have focused on twins and siblings (*see* Subheading 5.1) and include the Queensland Twin Imaging (QTIM) study, the Queensland Twin Adolescent Brain Project, the Vietnam Era Twin Study of Aging (VETSA) [28], and the Human Connectome Project (HCP) [29] (*see* Subheading 5.1). In addition, there are many more datasets available on neurological disorders, which may be explored via XNAT, LONI, or the Dementias Platform UK (DPUK), to name a few, PPMI (Parkinson's Progression Markers Initiative) [30], MEMENTO (deterMinants and Evolution of AlzheiMer's disEase aNd relaTed disOrder) [31], EPAD (European Prevention of Alzheimer's Dementia) [32], and ABIDE (Autism Brain Imaging Data Exchange) [33, 34].

3.2 Positron Emission Tomography (PET-MRI)

Positron emission tomography (PET) images are 3D images that highlight the concentration of a radioactive tracer administered to the patient. Here, we will focus on brain PET images, although other parts of the body may also be imaged. The different tracers allow to measure several aspects of brain metabolism (e.g., glucose) or spatial distribution of a molecule of interest (e.g., amyloid).

PET relies on the nuclear properties of radioactive materials that are injected in the patient intravenously. When the radioactive isotope disintegrates, it emits a photon that will be detected by the scanner. This signal is used to find the position of the emitted positrons which allow us to reconstruct the concentration map of the molecule we are tracing [35].

As for MRI, PET images are available as a series of DICOM files or a single NIfTI format. They are composed of voxels (3D pixel), and their size (e.g., 1 × 1 × 1 mm) corresponds to the image resolution. A BIDS extension has also been developed for positron emission tomography, in order to standardize data organization for research purposes.

PET is considered invasive due to the injection of the tracer, which results in very small risk of potential tissue damage. Overall, the quantity of radioactive isotope remains small enough to make it safe for most people, but this limits its widespread acquisition in research, especially on healthy subjects or in children. Moreover, PET requires to have a high-cost cyclotron to produce the radio-tracers nearby because the half-life of the radioisotopes is typically short (between a few minutes to few hours).

Several tracers are used for brain PET imaging, one of the most common ones being the ^{18}F -fluorodeoxyglucose (^{18}F -FDG). ^{18}F -FDG concentrates in areas that consume a lot of glucose and will thus highlight brain metabolism. In practice, ^{18}F -FDG PET

images are often used to study neurodegenerative disease by revealing hypometabolism that characterizes some dementia [36, 37]. Other diseases such as epilepsy and multiple sclerosis can be studied through this modality, but since it is not part of clinical routine, data are rare, and we are not aware of publicly available datasets.

In whole-body PET scans, ¹⁸F-FDG is used to detect tumors, which consumes a lot of glucose. However, the brain consumes a lot of glucose as part of its normal functioning, and brain tumors are not noticeable using this tracer. Instead, clinicians would use ¹¹C-choline that will also accumulate in the tumor area but is not specifically used by the brain otherwise. In addition to glycemic radiotracers, oxygen-15 is also used to measure blood flow in the brain, which is thought to be correlated with brain activity. In practice, this tracer is less used than ¹⁸F-FDG because of its very short half-life. Other tracers are used to show the spatial concentration of specific biomarkers: for instance, ¹⁸F-florbetapir (AV45), ¹⁸F-flutemetamol (Flute), Pittsburgh compound B (PiB), and ¹⁸F-florbetaben (FBB) are amyloid tracers used to highlight β-amyloid aggregation in the brain, which is a marker of Alzheimer's disease. Finally, ¹¹C-5-hydroxytryptamine (5-HT) neurotransmitter is used to expose the serotonergic transmitter system.

We have made a non-exhaustive list of publicly available datasets containing PET scans with different tracers. Most datasets focused on neurodegenerative disorders and also collected brain MRI (see previous section). The Alzheimer's Disease Neuroimaging Initiative (ADNI) is one of the largest datasets with PET images for Alzheimer's disease [11]. ADNI used F-FDG-PET as well as PET amyloid tracers: FBB, AV45, and PiB. The Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) only collected amyloid tracers of PET images: PiB, AV45, and Flute [12]. The Open Access Series of Imaging Studies v3 (OASIS3) includes PET imaging from three different tracers, PIB, AV45, and ¹⁸F-FDG [13].

In addition to those neurodegenerative datasets, PET is available in the Lundbeck Foundation Centre for Integrated Molecular Brain Imaging (CIMBI) database and biobank established in 2008 in Copenhagen, Denmark [38]. CIMBI shares structural MRI, PET, genetic, biochemical, and clinical data from 2000 persons (around 1600 healthy subjects and almost 400 patients with various pathologies). Tracer used for PET is the ¹¹C-5-HT which is relevant to study the serotonergic transmitter system. Applications to access the data can be made on their website by completing a form (*see Table 2*).

The ChiNese brain PET Template (CNPET) dataset has been developed by the Medical Imaging Research Group (<https://biomedimg-dlut-edu.cn/>), from Dalian University of Technology (China) [39]. The database contains 116 records of ¹⁸F-FDG-PET

from healthy patients, which has been used to make a Chinese population-specific statistical parametric mapping (SPM, i.e., average template used for PET processing). The data used to build the PET brain template have been released and are available on Neuro-Imaging Tools and Resources Collaboratory (NITRC, <https://www.nitrc.org/>) platform.

4 EEG/MEG

Electroencephalography (EEG) measures the electrical activity of the brain [40–42]. Signals are captured through sensors distributed over the scalp (noninvasive) or by directly placing the electrodes on the brain surface, a procedure that requires a surgical intervention [43]. This technique is characterized by its high temporal resolution, enabling the study of dynamic processes such as cognition or the diagnosis of conditions such as epilepsy. Yet, EEG signals are nonstationary and have a non-linear nature, which makes it difficult to get useful information directly in the time domain. Nonetheless, specific patterns can be extracted using advanced signal processing techniques.

Another technique that captures brain activity is magnetoencephalography (MEG). This technology maps the magnetic fields induced above the scalp surface. Similar to EEG, MEG provides high time resolution, but it is preferentially sensitive to tangential fields from superficial sources [44, 45]. This could be considered as an advantage, since magnetic fields are less sensitive to tissue conductivities, facilitating source localization. However, MEG instrumentation is more expensive and not portable [46, 47].

During signal recording, undesirable potential coming from sources other than the brain may alter the quality of the signals. These artifacts should be detected and removed in order to improve pattern recognition. Multiple methods could be applied depending on the artifact to be eliminated: re-referencing with common average reference (CAR), ICA decomposition to remove other physiological sources as eye movements or cardiac components, notch filter to get rid of power line noise, and pass-band filtering to keep the physiological rhythms of interest, among others [48–51]. Other spatial filters such as common spatial pattern (CSP) for channel selection or filter bank CSP (FBCSP) for band elimination are largely used in motor decoding [52, 53].

Other signal processing tools allow the user to extract features describing relevant information contained in the signals. Subsequently, those patterns may be used as input for a classification pipeline. The target features vary according to the condition under study. Generally, the domain of clinical diagnostics focuses either on event-related potentials (ERP) or on spectral content of the signal [54, 55]. The first refers to voltage fluctuations associated

with specific sensory stimuli (e.g., P300 wave) or task, like motor preparation and execution, covert mental states, or other cognitive processes. The amplitude, latency, and spatial location of the resulting waveform activity reveal the underlying mental state [56]. On the other side, spectral analysis refers to the computation of the energy distribution of the signals in the frequency domain. Most spectral estimates are based on Fourier transform; this is the case of non-parametric methods, such as Welch periodogram estimation, which based their computation on data windowing [57].

Another approach is to study the interactions across sources (inferring connection between two electrodes by means of temporal dependency between the registered signals), which is known as functional connectivity. Multiple connectivity estimators have been developed to quantify this interaction [58]. Through these functional interactions, complex network analysis can also be implemented, where sensors are modeled as nodes and connectivity interactions as links [59–61].

EEG and MEG are essential to evaluate several types of brain disorders. One of the most documented is epilepsy, based on seizure detection and prediction [62–64]. Other neurological conditions can be characterized like Alzheimer's disease, associated with changes in signal synchrony [65, 66]. Furthermore, motor task decoding in brain-computer interfaces (BCI) offers a promising tool in rehabilitation [67]. This type of data, from healthy to clinical cases, can be found on multiple open-access repositories, such as Zenodo (<https://zenodo.org>) or PhysioNet ((1)), as well as via collaborative projects such as the BNCI Horizon 2020 (<http://bnci-horizon-2020.eu>), which gathers a collection of BCI datasets (*see* Table 2). These repositories are also valuable in that they contribute to establishing harmonization procedures in processing and recording. All dataset-collected informed consent and data were anonymized to protect the participants' privacy. Moreover, regulations may vary from one country to another, which require, for example, studies to be approved by ethics committees. Additionally, licensing (that define copyrights of the dataset) must be considered depending on the intended use of the open-access datasets.

Data come in different formats according to the acquisition system or the preprocessing software. The most common formats for EEG are .edf, .gdf, .eeg, .csv, or .mat files. For MEG, it is very often .fif and .bin (*see* Table 1). The different formats can create challenges when working with multiple datasets. Luckily, some tools have been developed to handle this problem, for example, FieldTrip [68] or Brainstorm [69] implemented in MATLAB, or the Python modules mne [70] and moabb [71]. Of note, these tools also contain sets of algorithms and utility functions for analysis and visualization.

5 Genetics

5.1 Twin Samples

Twins provide a powerful method to estimate the importance of genetic and environmental influences on variation in complex traits. Monozygotic (MZ, aka identical) twins develop from a single zygote and are (nearly) genetically identical. In contrast, dizygotic (DZ, aka fraternal) twins develop from two zygotes and are, on average, no more genetically related than non-twin siblings. In the classical twin design, the degree of similarity between MZ and DZ twin pairs on a measured trait reveals the importance of genetic or environmental influences on variation in the trait. Twin studies often collect several different data types, including brain MRI scans, assessments of cognition and behavior, self-reported measures of mental health and wellbeing, as well as biological samples (e.g., saliva, blood, hair, urine). Datasets derived from twin studies are text-based and include phenotypic data and background variables (e.g., individual and family IDs, sex, zygosity, age). Notably, the correlated nature of twin data (i.e., the non-independence of participants) should be considered during analysis as it may violate statistical test assumptions [72, 73].

Raw data is typically stored locally by the data owner, with de-identified data available upon request. In larger studies, data is stored and distributed through online repositories. Recently, the sharing of publicly available de-identified data with accompanying publications has become commonplace.

Several extensive twin studies combine imaging, behavioral, or biological data (*see* Table 2). These studies cover the whole life span (STR) as well as specific age periods, for example, children/adolescents (QTAB), young (QTIM, HCP-YA), middle-aged (VETSA), or older (OATS) adults.

The Swedish Twin Registry (STR) was established in the late 1950s with the primary aim to explore the effect of environmental factors (e.g., smoking and alcohol) on disorders [74]. Data were first collected through questionnaires and interviews with the twins and their parents. Later, the STR incorporated data from biobanks, clinical blood chemistry assessments, genotyping, health checkups, and linkages to various Swedish national population and health registers [74]. The STR is now one of the largest twin registers in the world [75] with information on more than 87,000 twin pairs (<https://ki.se/en/research/the-swedish-twin-registry>). It has been used extensively for the research of health and illness, including various neurological disorders, including dementia [76], Parkinson's disease [77], and motor neuron disease [78].

The Queensland Twin Adolescent Brain (QTAB, 2015–present) was enabled through funding from the Australian National Health and Medical Research Council (NHMRC). It focuses on the period of late childhood/early adolescence, with brain imaging,

cognition, mental health, and social behavior data collected over two waves (age 9–14 years at baseline, $N = 427$). A primary objective is to chart brain changes and the emergence of depressive symptoms throughout adolescence. Biological samples (blood, saliva), sleep (self-report), and motor activity measures (see Section 8) were also collected. Data is available from the project owners upon request.

The Queensland Twin IMaging (QTIM, 2007–2012) study, funded through the National Institutes of Health (NIH) and NHMRC, was a collaborative project between researchers from QIMR Berghofer Medical Research Institute, the University of Queensland, and the University of Southern California, Los Angeles. Brain imaging was collected in a large genetically informative population sample of young adults (18–30 years, $N > 1200$) for whom a range of behavioral traits, including cognitive function, were already characterized (as a component of the Brisbane Adolescent Twin Study, QIMR Berghofer Medical Research Institute [79]). Notably, the dataset includes a test–retest neuroimaging subsample ($n = 75$) to estimate measurement reliability. Data is available from the project owners upon request.

The Human Connectome Project Young Adult (HCP-YA, 2010–2015) study, funded by the NIH, is based at Washington University, University of Minnesota, and Oxford University. Investigators spent 2 years developing state-of-the-art imaging methods [29] before collecting high-quality neuroimaging, behavioral, and genotype data in ~1200 healthy young adult twins and non-twin siblings (22–35 years). HCP-YA data has been used widely in twin-based analyses, examining genetic influences on network connectivity [80], white matter integrity [81], and cortical surface area/thickness [82]. Open-access HCP-YA data is available from the Connectome Coordination Facility following registration (<https://db.humanconnectome.org>), with additional data use terms applicable for restricted data (e.g., family structure, age by year, handedness).

The Vietnam Era Twin Study of Aging (VETSA, 2003–present), funded by the NIH, started as a study of cognitive and brain aging but has since pivoted to the early identification of risk factors for mild cognitive impairment and Alzheimer’s disease [28]. In addition to neuroimaging and cognitive data, the VETSA study includes health, psychosocial, and neuroendocrine data collected across three waves (baseline mean age 56 years, follow-up waves every 5–6 years) [83]. VETSA data is available following registration (<https://medschool.ucsd.edu/som/psychiatry/research/VETSA/Researchers/Pages/default.aspx>).

The Older Australian Twins Study (OATS, 2007–present) [84], funded by the NHMRC and Australian Research Council, is a longitudinal study of genetic and environmental contributions to brain aging and dementia. The project includes neuroimaging and

cognitive data collected across four waves (baseline mean age 71 years, follow-up waves every 2 years). OATS was expanded in wave 2 to include positron emission tomography (PET) scans to investigate the deposition of amyloid plaques in the brain. Data is available from the project owners upon request.

There is a wealth of twin studies worldwide in addition to those mentioned here (see [85] for an overview). Foremost is the Netherlands Twin Registry [86], a substantial data resource with dedicated projects investigating neuropsychological, biomarker, and behavioral traits. In addition, several extensive family/pedigree imaging studies exist, including the Genetics of Brain Structure and Function study [87] and the Diabetes Heart Study-Mind Cohort [88]. Further, the previously mentioned ABCD study [89] includes embedded twin subsamples.

Twin datasets have been used to estimate the heritability (the proportion of observed variance in a phenotype attributed to genetic variance) of phenotypes derived through machine learning, such as brain aging [90–92] and brain network connectivity [93]. Further, machine learning models have been trained to discriminate between MZ and DZ twins based on dynamic functional connectivity [94] and psychological measures [95]. In addition, machine learning has been used to predict co-twin pairs based on functional connectivity data [96].

5.2 Molecular Genetics

5.2.1 The UK Biobank

The UK Biobank (UKB) is one of the largest population-based cohorts, comprising nearly half a million adult participants (aged over 40 years at the time of recruitment), recruited across over 20 assessment centers in the UK. The UKB resource is accessible to the research community through application (<https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access>) and, as of the end of 2021, counted more than 28,000 registered approved researchers worldwide. In 2021, UKB launched a cloud-based Research Analysis Platform (RAP), which provides computational tools for data visualization and analysis, thereby aiming to democratize access for researchers lacking such infrastructure. The associated fees for using the UKB resource include the yearly tier-based access fees, depending on the type of data accessed, as well as the cost of running the analyses and storing the generated data, while the storage of the UKB dataset itself is provided free of charge. Certain emerging datasets (e.g., whole exome and genome sequences) will be only available for analysis through the platform, both due the enormous size and tighter regulation around those datasets. Upon publication, researchers are required to return their results, including the methodology and any essential derived data fields, back to the UKB, which are subsequently incorporated into the resource in order to promote reproducible research.

The cohort is deeply phenotyped with thousands of traits measured across multiple assessments. The initial assessment visit took place from 2006 to 2010, where ~502,000 participants consented to participate (each keeping the right to withdraw their consent and be removed from the study at any time), completed the interview, filled questionnaires, underwent multiple measurements, and donated blood urine and saliva samples (see <https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/Reception.pdf>).

The first repeat assessment was conducted in 2012–2013 and included approximately 120,000 participants. Next, the participants were invited to attend the imaging visits: the initial (2014+) and the first repeat imaging visit (2019+). So far, 50,000 initial imaging visits have been conducted, with a target to image 100,000 participants (10,000 repeat). The imaging data includes brain [14, 97], heart [98], and abdominal MRI scans [99], with both bulk images and image-derived measures available for analysis, as well as retinal OCT images, whole body MRI, and carotid ultrasound [100]. Finally, follow-up information from the linked health and medical records is regularly collected and updated in the resource, including data for COVID-19 research. The showcase of the available anonymous summary information is available at <https://biobank.ndph.ox.ac.uk/showcase/>.

The interim release of the genotyping data comprised ~150,000 samples and was released in 2015, followed by the full release of 488,000 genotypes in the middle of 2017. The available genotype data included variant calls from UK BiLEVE and UK Biobank Axiom arrays (autosomes, sex chromosomes, and mitochondrial DNA) as well as phased haplotype values and imputation to a combined panel of Haplotype Reference Consortium (HRC) and the merged UK10K and 1000 Genomes phase 3 reference panels [101], also known as v2 release. Subsequently, the v2 imputation was replaced by imputation to HRC and UK10K haplotype resource only (v3), after a problem was discovered for the set imputed to UK10K + 1000 Genomes panel (<https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100319>). The genotypes of approximately 3% of the participants remained not assayed due to DNA processing issues. To note, ~50,000 individuals included in the interim genotype release were involved in the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) project, and their genotypes were assayed on a different but very closely related array than the rest of the participants (https://biobank.ctsu.ox.ac.uk/crystal/ukb/docs/genotyping_qc.pdf). The UK BiLEVE focused on genetics of respiratory health, and the participants were selected based on lung function and smoking behavior [102].

Whole exome sequencing (WES) and whole genome sequencing (WGS) have been funded through the collaboration between the UK Biobank and biotechnology companies Regeneron and GlaxoSmithKline (GSK). The first UKB release of WES data

included 50,000 participants, prioritized based on the availability of MRI data, baseline measurements, and linked hospital and primary care records and enriched in patients diagnosed with asthma [103]. Recently (November 2021), the new data release included $N = 200,000$ WGS and $N = 450,000$ WES [104]. WGS for the remaining participants is currently underway. For all the past and future timelines, see https://biobank.ctsu.ox.ac.uk/showcase/exinfo.cgi?src=timelines_all.

Most of the UKB participants reported their ethnic background as White British/Irish or any other white background (~94%), which was coherent with the observed genetic ancestries [101]. For example, the ancestries identified from genetic markers showed a predominant European ancestry ($N \sim 464,000$), followed by South Asian (~12,000), African (~9000), and East Asian ancestry (~2500) [105]. As a population-based cohort, the UKB mostly comprises unrelated participants. While the pedigree information was not collected as a part of assessment, the genetic analysis has identified approximately 100,000 pairs of close relatives (third degree or closer, including 22,000 sibling pairs and 6000 parent-offspring pairs) [101]. This amount of relatedness is, however, larger than expected for a random sample from a population and reflects a participation bias toward the relatives of the participants. Moreover, the UKB sample is, on average, healthier, more educated, and less deprived than the general UK population [2].

5.3 Genetic Consortia

5.3.1 ENIGMA Consortium

The Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) consortium was formed in 2009 with the goal of conducting large-scale neuroimaging genetic studies of human brain structure, function, and disease [27]. Currently, more than 2000 scientists from 400 institutions around the world with neuroimaging (including structural and functional MRI) and electroencephalography (EEG) data have joined the consortium and formed 50 working groups that focus on different psychiatric and neurological disorders as well as healthy variation, method development, and genomics [27].

To date, the ENIGMA Genetics Working Group (for an overview, see [106]) have conducted genome-wide association meta-analyses for hippocampal and intracranial volume [107–109], subcortical volume [110, 111], and cortical surface area and thickness [112]. The ENIGMA Genetics Working Group provides researchers imaging and genetic protocols to enable each group to conduct their own association analyses before contributing summary statistics to the meta-analysis. While these genome-wide association studies have focused on structural phenotypes and the analysis of common single nucleotide polymorphisms (SNPs), the ENIGMA EEG Working Group have recently conducted a genome-wide association meta-analysis for oscillatory brain activity [113], and the ENIGMA Copy Number Variant (CNV) Working Group,

which formed in 2015, is currently investigating the impact of rare CNVs beyond the 22q11.2 locus on cognitive, neurodevelopmental, and neuropsychiatric traits [114].

The sample sizes of the ENIGMA Genetics and CNV Working Groups continuously increase as new cohorts with MRI and genetic data join the consortium. As of 2020, the CNV Working Group sample comprises of 38 ENIGMA cohorts [114], while the latest Genetics Working Group genome-wide association meta-analysis [112] consisted of a discovery sample of 49 ENIGMA cohorts and the UK Biobank ($N = 33,992$ individuals of European ancestry), a replication sample of 2 European ancestry cohorts ($N = 14,729$ participants), and 8 ENIGMA cohorts of non-European ancestry ($N = 2994$ participants). This meta-analysis identified 199 genome-wide significant variants that were associated with either the surface area or thickness of the whole human cortex and 34 cortical regions with known functional specializations. They also found evidence that the genetic variants that influence brain structure also influence brain function, such as general cognitive function, Parkinson's disease, depression, neuroticism, ADHD, and insomnia [112].

Importantly, all imaging, EEG, and genetic (imputation and association analysis) protocols are freely available from the ENIGMA website (<http://enigma.ini.usc.edu/>). However, to access the summary statistics for each published genome-wide association meta-analysis, researchers need to complete an online Data Access Request Form (<http://enigma.ini.usc.edu/research/download-enigma-gwas-results/>). If a researcher wants to propose new genetic analyses that cannot be conducted with these publicly available summary statistics, they need to become a member of ENIGMA. Researchers can join the consortium by (a) contributing a cohort with MRI and genetic data, (b) collaborating with another research group that does have MRI and genetic data, or (c) contributing their expertise in genomic or methodological areas that are inadequately addressed by other consortium members. Of note, since storage of the MRI and genetic data is not centralized, each ENIGMA cohort can choose to contribute or not to new proposed analyses.

5.3.2 The Psychiatric Genomics Consortium (PGC)

The Psychiatric Genomics Consortium (PGC) began in 2007. The central idea of the PGC is to use a global cooperative network to advance genetic discovery in psychiatric disorders in order to identify biologically, clinically, and therapeutically meaningful insights. To date, the PGC is one of the largest, most innovative, and productive consortia in the history of psychiatry. The Consortium now consists of workgroups on 11 major psychiatric disorders, a Cross-Disorder Workgroup, and a Copy-Number Variant Workgroup. In addition, the PGC provides centralized support to the PGC researchers with a Statistical Analysis Group, Data Access

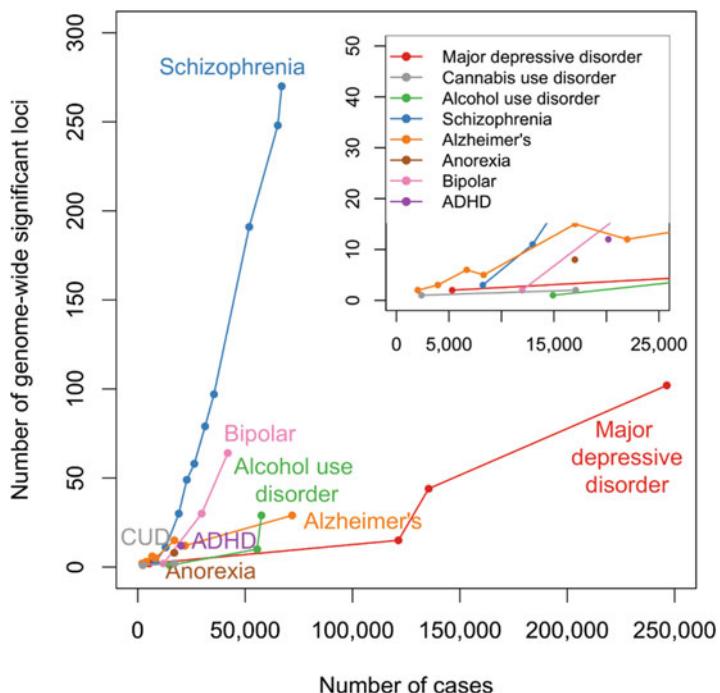


Fig. 1 PGC discoveries over time

Committee, and Dissemination and Outreach Committee. To increase ancestral diversity, the Consortium established the Cross-Population Workgroup in 2017 for outreach and developing/deploying trans-ancestry analysis methods [115]. The Consortium outreach expands ancestry diversity by adding non-European cases and controls. The PGC continues to unify the field and attract outstanding scientists to its central mission (800+ investigators from 150+ institutions in 40+ countries). PGC work has led to 320 papers, many in high-profile journals (*Nature* 3, *Cell* 5, *Science* 2, *Nat Genet* 27, *Nat Neurosci* 9, *Mol Psych* 37, *Biol Psych* 25, *JAMA Psych* 12). The full results from all PGC papers are freely available, and the findings have fueled analyses by non-PGC investigators (sample sizes and findings for eight major psychiatric disorders are summarized in Fig. 1)

Computation and data warehousing for the PGC are non-trivial. The PGC uses the Netherlands “LISA” computing cluster. LISA compute cluster in Amsterdam which is used for most analyses (occasional analyses are done on other clusters, but 90% of PGC computation is done on LISA). The core software is the RICOPILI data analytic pipeline [116]. This pipeline has explicit written protocols for uploading data to the cluster in the Netherlands that one uses for quality control, imputation, analysis, meta-analysis, and bioinformatics. The actual mega-analyses are conducted by PGC analysts under the direction of a senior statistical geneticist, geneticist, or highly experienced analyst.

The PGC has a proven commitment to open-source, rapid progress science. All PGC results are made freely available as soon as a primary paper is accepted (GWAS summary statistics available at <https://www.med.unc.edu/pgc/download-results/>). The researchers can obtain access to the individual-level data either through controlled-access repositories (e.g., the Database of Genotypes and Phenotypes, dbGaP, or the European Genome-phenome Archive) or via the PGC streamlined process for secondary data analyses (<https://www.med.unc.edu/pgc/shared-methods/data-access-portal/>) [117].

PGC analyses have always been characterized by exceptional rigor and transparency. PGC analysts will enhance this by publishing markdown notebooks for all papers on the PGC GitHub site (<https://github.com/psychiatric-genomics-consortium>) to enable precise reproduction of all analyses (containing code, documentation of QC decisions, analyses, etc.).

5.4 Exome and Whole Genome Sequencing: Trans-Omics for Precision Medicine (TOPMed)

The Trans-Omics for Precision Medicine (TOPMed) program, sponsored by the National Institutes of Health (NIH) National Heart, Lung, and Blood Institute (<https://topmed.nhlbi.nih.gov>), is part of a broader Precision Medicine Initiative, which aims to provide disease treatments tailored to an individual's unique genes and environment. TOPMed contributes to this Initiative through the integration of whole genome sequencing (WGS) and other omics data. The initial phases of the program focused on whole genome sequencing of individuals with rich phenotypic data and diverse backgrounds. The WGS of the TOPMed samples was performed over multiple studies, years, and sequencing centers [118, 119]. Available data are processed periodically to produce genotype data “freezes.” Individual-level data is accessible to researchers with an approved dbGaP data access request (<https://topmed.nhlbi.nih.gov/data-sets>), via Google and Amazon cloud services. More information about data availability and how to access it can be found on the dataset page (<https://topmed.nhlbi.nih.gov/data-sets>).

As of September 2021, TOPMed consists of ~180 K participants from >85 different studies with varying designs. Prospective cohorts provide large numbers of disease risk factors, subclinical disease measures, and incident disease cases; case-control studies provide improved power to detect rare variant effects. Most of the TOPMed studies focus on HLBS (heart, lung, blood, and sleep) phenotypes, which leads to 62 K (~35%) participants with heart phenotype, 50 K (~28%) with lung data, 19 K (~11%) with blood, 4 K (~2%) with sleep, and 43 K (~24%) for multi-phenotype cohort studies. TOPMed participants' diversity is assessed using a combination of self-identified or ascriptive race/ethnicity categories and observed genetics. Currently, 60% of the 180 K sequenced

participants are of non-European ancestry (i.e., 29% African ancestry, 19% Hispanic/Latino, 8% Asian ancestry, 4% other/multiple/unknown).

Whole genome sequencing is performed by several sequencing centers to a median depth of 30 \times using DNA from blood, PCR-free library construction, and Illumina HiSeq X technology (<https://topmed.ncbi.nih.gov/group/sequencing-centers>). Randomly selected samples from freeze 8 were used for whole exome sequence using Illumina v4 HiSeq 2500 at an average 36.4 \times depth. A trained machine learning algorithm with known variants and Mendelian inconsistent variants is applied by the Informatics Research Centre for joint genotype calling across all samples to produce genotype data “freezes” (<https://topmed.ncbi.nih.gov/group/irc>). In TOPMed data freeze 8 ($N \sim 180$ K) (<https://topmed.ncbi.nih.gov/data-sets>), variant discovery identified 811 million single nucleotide variants and 66 million short insertion/deletion variants. In the latest data freeze 9 (<https://topmed.ncbi.nih.gov/data-sets>), variant discovery was initially made on ~206 K samples including data from Centers for Common Disease Genomics (CCDG). This data was re-subset to ~158,470 TOPMed samples plus 2504 from 1000 Genomes samples were used for variant re-discovery. Then, a total of 781 million single nucleotide variants and 62 million short insertion/deletion variants were identified and passed variant quality controls. These variant counts in freeze 9 are slightly smaller than that of freeze 8 due to monomorphic sites in TOPMed samples. A series of data freezes is being made available to the scientific community as genotypes and phenotypes via dbGaP (<https://www.ncbi.nlm.nih.gov/gap/>); read alignments are available via the Sequence Read Archive (SRA) and variant summary information via the Bravo variant server (<https://bravo.sph.umich.edu/freeze8/hg38/>) and dbSNP (<https://www.ncbi.nlm.nih.gov/snp/>).

TOPMed studies provide unique opportunities for exploring the contributions of rare and noncoding sequence variants to phenotypic variation. For instance, [119] used 53,831 samples from freeze 5 (<https://topmed.ncbi.nih.gov/data-sets>) to investigate the role of rare variants into mutational processes and recent human evolutionary history. The recent TOPMed freeze 8 were used (together with WGS from the UK Biobank) to assess effect size of casual variants for gene expression using 72 K African American and ~298 K European American [120]. Similarly, a large set of multi-ethnic samples from freeze 5, 8, and 9 were used to develop comprehensive tools such as the STAAR and SCANG pipelines, which are used to identify noncoding rare variants [121] and to build predictive models for protein abundances [122] and discovery of causal genetic variants for different phenotypes [123, 124]. Overall, the Trans-Omics for Precision Medicine (TOPMed) program has the potential to help in improving

diagnosis, treatment, and prevention of major diseases by adding WGS and other “omics” data to existing studies with deep phenotyping.

6 Genomics

6.1 Methylation

DNA methylation (DNAm) is a covalent molecular modification by which methyl groups (CH_3) are added to the DNA. In vertebrates—and eukaryotes in general—the most common methylation modification occurs at the fifth carbon of the pyrimidine ring (5mC) at cytosine–guanine dinucleotides (CpG). Most bulk genomic methylation patterns are stable across cell types and throughout life, changing only in localized contexts, for example, due to disease-associated processes.

There are numerous ways of measuring DNAm at a genome-wide level, with bisulfite conversion-based methods being the most popular in the field of epidemiological epigenetics. These methods consist of bisulfite-induced modifications of genomic DNA, which results in unmodified cytosine nucleotides being converted to uracil, while 5mC remain unaffected. Of all these bisulfite conversion-based technologies—including sequencing-based methods—hybridization arrays are the most widely used, primarily due to their low cost and high-throughput nature.

The current Illumina Infinium® HumanMethylation450 (or 450 K) and Illumina Infinium® HumanMethylation850 (or EPIC) arrays assess around 450,000 and 850,000 methylation sites across the genome, respectively, covering 96% of the CpG islands (i.e., genomic regions with high CpG frequency), 92% of the CpG islands’ shores [125, 126] (<2 kb flanking CpG Islands), and 86% of the CpG islands’ shelves (<2 kb flanking outward from a CpG shore), which have been shown to be more dynamic than CpG islands [127]. Although most current studies have used the 450 K array [128], the EPIC array covers >90% of the 450 K sites plus additional CpG sites in the enhancer regions identified by the ENCODE and FANTOM5 projects [129].

After probe hybridization and extension steps, the array is scanned, and the intensities of the unmethylated and methylated bead types are measured. DNAm values are then represented by the ratio of the intensity of the methylated bead type to the combined locus intensity. These are known as *beta* (β) values and are continuous variables between 0 and 1 (Equation 1), although a value of 1 is impossible to achieve in practice, due to the addition of a stabilizing α offset (to handle low-intensity signals):

Equation 1 DNA methylation β values as measured by the Illumina Infinium® methylation arrays M = methylated intensity, U = unmethylated intensity, α = arbitrary offset to handle signals with low readings (usually 100)

$$\beta = \frac{M}{M + U + \alpha} \quad (1)$$

These raw intensities are then stored in binary IDAT files (one for each of the red and green channels). The bulk of each file consists of four fields: the ID of each bead type on the array, the mean and standard deviation of their intensities, and the number of beads of each type, generated per sample. This raw data format allows for flexible use, including differing preprocessing strategies [130]. However, these files are usually not readily available in public repositories (e.g., Gene Expression Omnibus [131] or GEO), due to their large size. For example, a compressed .tar file of IDATs for a sample size of around 700 individuals, measured with EPIC arrays, is about 10 Gb. Instead, researchers usually upload the processed DNAm β values (following normalization) as compressed .txt or .csv files with columns representing samples and rows the measured *loci*. This can be a problem for reproducibility, as different research groups tend to prefer their own preprocessing or normalization methods—and there are many [132]! On this note, there has been a recent push in the field, for standardization of DNAm array preprocessing pipelines, including the user-friendly *Meffil* pipeline [133].

Reproducibility and interpretation of DNAm studies are subject to additional factors outside of data processing methods. For comparison, genetic data is (mostly) germline determined and can be assumed to be randomly assigned with respect to characteristics of individuals. Thus, a case-control (or cross-sectional) design has an inference of association through causality and can convey information of liability to disease. This contrasts with DNAm data which is a reversible process influenced by a large range of biological, technical, and environmental factors (e.g., medication and complications of the disease itself) and is thus more susceptible to spurious cryptic association or reverse causation [134, 135]. DNAm studies will therefore benefit from longitudinal designs, both for biomarker discovery and mechanistic insights [134, 136].

Reed et al. [137] provide one good example of this. Briefly, the authors generated a DNAm score for body mass index (BMI) within the ARIES subsample of the Avon Longitudinal Study of Parents and Children birth cohort (ALSPAC), using effect sizes of 135 CpG sites from a published meta-analysis of DNAm and BMI [138]. Using multiple time points for matched mothers and children using linear and cross-lagged models to explore the causal relationship between phenotypic BMI and the DNAm scores, they

found a strong linear association within time points [137]. However, when testing for temporal associations, DNAm scores at earlier time points showed no association with future BMI, indicating that a DNAm score generated from a reference cross-sectional study performs better as a biomarker of extant BMI, but poorly as a predictor for future BMI.

In Table 2, we have compiled a list of the largest and/or most used DNAm array datasets—including the Genetics of DNA Methylation Consortium (goDMC), an international collaboration of human epidemiological studies that comprises >30,000 study participants with genetic and DNAm array data [139]. These samples are usually integrated in larger genetic/epidemiological studies, except for perhaps the NIH Roadmap Epigenomics Mapping Consortium [140], which was launched with the goal of producing a public resource of human epigenomic data to catalyze basic biology and disease-oriented research, and the BLUEPRINT project [141, 142], which aims to generate at least 100 reference epigenomes of distinct types of hematopoietic cells from healthy individuals and of their malignant leukemic counterparts. Lastly, in contrast to genetic data, the de-identified DNAm data—either raw or preprocessed—is typically open access in public repositories such as GEO [131], or dbGAP [143], or the web portals provided by the respective projects. However, access to accompanying phenotypic data may require additional approval by the managing committees of each individual project.

6.2 Gene Expression Data: GTEx

Launched in 2010, the Genotype-Tissue Expression (GTEx) project is an ongoing effort that aims to characterize the genetic determinants of tissue-specific gene expression [144]. It is a resource database available to the scientific community, which is comprised of multi-tissue RNA sequencing (RNA-seq: gene expression) and whole genome sequence (WGS) data collected in 17,382 samples across 54 tissue types from 948 postmortem donors (version 8 release). Sample size per tissue ranges from $n = 4$ in kidney (medulla) to $n = 803$ in skeletal muscle. The majority of donors are of European ancestry (84.6%) and male (67.1%) with ages ranging from 20–70 years old. The primary cause of death for donors 20–39 years old was traumatic injury (46.4%) and heart disease for donors 60–70 years (40.9%).

Data is constantly being added to the database using sample data from the GTEx Biobank. For example, recent efforts have focused on gene expression profiling at the single-cell level to achieve a higher resolution understanding of tissue-specific gene expression and within tissue heterogeneity. As a result, single-cell RNA-seq (scRNA-seq) data was generated in 8 tissues from 25 archived, frozen tissue samples collected on 16 donors. Further, the Developmental Genotype-Tissue Expression (dGTEx) project (<https://dgtex.org/>) is a relatively new extension of GTEx that was

launched in 2021 that aims to understand the role of gene expression at four developmental time points: postnatal (0–2 years of age), early childhood (2–8 years of age), pre-pubertal (8–12.5 years of age), and post-pubertal (12.5–18 years of age). It is expected that molecular profiling (including WGS, bulk RNA-seq, and, for a subset of samples, scRNA-seq) will be performed on 120 relatively healthy donors (approximately 30 donors per age group) in 30 tissues. Data from this study would provide, for example, a baseline for gene expression patterns in normal development for comparison against individuals with disease.

GTEx provides extensive documentation on sample collection, laboratory protocols, quality control and standardization, and analytical methods on their website (<https://gtexportal.org/home/>). This allows for replication of their protocols and procedures in other cohorts to aid in study design and for researchers to further interrogate the GTEx data to answer more specific scientific questions. Processed individual-level gene expression data is made freely available on the GTEx website for download, while controlled access to individual-level raw genotype and RNA sequencing data are available on the AnVIL repository following approval via the National Center for Biotechnology Information's database of Genotypes and Phenotypes (dbGAP, dbGaP accession phs000424), a data archive website that stores and distributes data and results investigating the relationship between genotype and phenotype (<https://www.ncbi.nlm.nih.gov/gap/>). Clinical data collected for each donor is categorized into donor-level (demographics, medication use, medical history, laboratory test results, death circumstances, etc.) and sample-level (tissue type, ischemic time, batch ID, etc.) data and is also available through dbGAP.

Over the many years, data from the GTEx project has provided unprecedented insight into the role genetic variation plays in regulating gene expression and its contribution to complex trait and disease variation in the population. The latest version 8 release from GTEx comes with a comprehensive catalogue of variants associated with gene expression, or eQTLs (expression quantitative trait loci), across 49 tissues or cell lines (derived from 15,201 samples and 838 donors) (GTEx Consortium, 2020). This analysis has demonstrated that gene expression is a highly heritable trait, with millions of genetic variants affecting the expression of thousands of genes across the genome. These pairwise gene variant associations can be classified as either *cis*- or *trans*-eQTLs, which describes proximal (i.e., within a predefined window of the target gene) or distal (i.e., beyond the predefined window or on a different chromosome from the target gene) genetic control, respectively. Indeed, it has been shown that 94.7% of all protein-coding genes have at least one *cis*-eQTL. In addition, 43% of genetic variants (minor allele frequency > 1%) have been found to affect gene expression in at least one tissue, and the majority of *cis*-eQTLs appear to be shared

across the sexes and ancestries (GTEx Consortium, 2020). Relatively few *trans*-eQTLs have been identified due to limitations in sample sizes; however, these typically affect gene expression in one or very few tissues, with about a third of *trans*-eQTLs mediated by *cis*-eQTLs [144]. Importantly, GTEx provides full eQTL summary statistics for download and an interactive portal (<https://gtexportal.org/home/>) for quick searches. As most trait-associated loci identified in genome-wide association studies (GWAS) are in noncoding regions of the genome, the eQTL data generated by GTEx has been leveraged to provide insight into the genetic and molecular mechanisms that underlie complex traits and diseases. Indeed, GWAS trait-associated variants are enriched for *cis*-eQTLs, and genetic variants that affect multiple genes in multiple tissues are found to also affect many complex traits (GTEx Consortium, 2020). This indicates that *cis*-eQTLs have a high degree of pleiotropy and exert their effect on complex traits and diseases by regulating proximal gene expression.

In addition to the comprehensive catalogue of multi-tissue eQTLs to understand gene regulation, additional flagship GTEx studies include understanding sex-biased gene expression across tissues [145], functional rare genetic variation [146], cell type-specific gene regulation [147], and predictors of telomere length across tissues [148].

The extensive publicly available data generated by the GTEx project is a valuable resource to the scientific community and will allow for further data interrogation for many years to come.

7 Electronic Health Records

7.1 Clinical Data Warehouse: Example from the Parisian Hospitals (APHP)

Clinical data warehouses (CDW) gather electronic health records (EHR), which can gather demographic data, results from biological tests, prescribed medications, and images acquired in clinical routine, sometimes for millions of patients from multiple sites. CDW can allow for large-scale epidemiological studies, but they may also be used to train and/or validate machine learning (ML) and deep learning (DL) algorithms in a clinical context. For example, several computer-aided diagnosis tools have been developed for the classification of neurodegenerative diseases. One of their main limitations is that they are typically trained and validated using research data or on a limited number of clinical images [149–154]. It is still unclear how these algorithms would perform on large clinical dataset, which would include participants with multiple diagnoses and more generally heterogeneous data (e.g., multiple scanners, hospitals, populations).

One of the first CDW in France was launched in 2017 by the AP-HP (Assistance Publique – Hôpitaux de Paris), which gathers most of the Parisian hospitals [155]. They obtained the

authorization of the CNIL (Commission Nationale de l'Informatique et des Libertés, the French regulatory body for data collection and management) to share data for research purposes. The aim is to develop decision support algorithms, to support clinical trials, and to promote multicenter studies. The AP-HP CDW keeps patients updated about the different research projects through a portal (as authorized by CNIL), but, according to French regulation, active consent was not required as these data were acquired as part of the routine clinical care of the patients.

Accessing the data is possible with the following procedure. A detailed project must be submitted to the Scientific and Ethics Board of the AP-HP. If the project holders are external to the AP-HP, they have to sign a contract with the Clinical Research and Innovation Board (Direction de la Recherche Clinique et de l'Innovation). Once the project is approved, data are extracted and pseudo-anonymized by the research team of the AP-HP. Data are then made available in a specific workstation via the Big Data Platform, which is internal to the AP-HP. The Big Data Platform supports several research environments (e.g., JupyterLab Environment, R, MATLAB) and provides computational power (CPUs and GPUs) to analyze the data.

An example of the research possible using such CDW is the APPRIMAGE project, led by the ARAMIS team at the Paris Brain Institute. The project was approved by the Scientific and Ethics Board of the AP-HP in 2018. It aims to develop or validate algorithms that predict neurodegenerative diseases from structural brain MRI, using a very large-scale clinical dataset. The dataset provided by the AP-HP gathers all T1w brain MRI of patients aged more than 18 years old, collected since 1980. It therefore consists of around 130,000 patients and 200,000 MRI which were made available via the Big Data Platform of the AP-HP. Of note, clinical data was available for only 30% of the imaged participants (>30,000 patients) as it relies on the ORBIS Clinical Information System (Agfa HealthCare), installed more recently in the hospitals. The sheer size of the data poses obvious computational challenges, but other difficulties include harmonizing clinical reports collected in the different hospitals or handling the general heterogeneity of the data (e.g., hospitals, acquisition software, populations). To tackle this issue, we have developed a pipeline for the quality control of the MR images [156].

7.2 Swedish National Registries

In Sweden, a unique 10-digit personal identification number has been assigned to each individual at birth or migration since 1947, which allows linkages across different Swedish population and health registers with almost 100% coverage [157]. The Swedish Total Population Register (TPR) was established in 1968 and is maintained by Statistics Sweden to obtain data on major life events, such as birth, vital status, migration, and civil status [158]. TPR is a

key source to provide basic information in medical and social research in Sweden. The Swedish Population and Housing Censuses (1960–1990) and the Swedish Longitudinal Integrated Database for Health Insurance and Labour Market Studies (Swedish acronym LISA) (since 1990) provide information on demographic and socioeconomic status for the Swedish population, including the highest attained educational level and household income [159]. The Swedish Multi-Generation Register (MGR) provides information on familial links for individuals born since 1932 onward in Sweden [160], which makes it possible to perform family studies to investigate familial risk of different health outcomes and control for familial confounding when needed.

The Swedish National Patient Register (NPR) is a valuable source for medical research, which has since 1964 collected data on inpatient care (nationwide coverage since 1987) and outpatient care (more than 85% of the entire country since 2001) [161]. Diagnoses are according to the Swedish revisions of the International Classification of Disease codes (ICD codes). The positive predictive value of the diagnoses is high, ranging from 85% to 95%, in NPR [161]. NPR has been used in studies of different diseases including many neurological disorders such as Alzheimer's disease [162], Parkinson's disease [163], and amyotrophic lateral sclerosis [164]. The Swedish Cancer Register (SCR) has been used extensively in Swedish cancer research, especially cancer epidemiology. SCR was established in 1958 and includes data on all newly diagnosed malignant and benign tumors, including different kinds of brain tumors [165, 166]. The Swedish Medical Birth Register (MBR) was established in 1973 and contains information on almost all deliveries (from prenatal to postnatal) in Sweden [167]. MBR has contributed mainly to the reproductive epidemiologic research in Sweden and has also been used in epidemiological studies of diseases later in life including different neurological disorders [168, 169]. The Swedish Causes of Death Register (CDR) includes information on virtually all deaths in Sweden since 1952 [170] and has been used to identify various causes of death in medical research, including deaths due to neurological disorders [171]. The Swedish Prescribed Drug Register (PDR) was founded in July 2005 and provides information on all prescription drugs dispensed from pharmacies in Sweden [172, 173]. PDR has been used to study patterns of use as well as consequences of medication use, including memantine [174] and dopaminergic anti-Parkinson drug [175].

In addition to these general health registers, there are also hundreds of disease quality registers that are used for patient care and research in Sweden. For instance, the Swedish Dementia Registry (SDR) was established in 2007 to achieve high quality of diagnostics and care for patients with dementia [176]. The Swedish Neuro-Register (SNR) was founded in 2001 (web-based since

2004, originally named as the Swedish Multiple Sclerosis Quality Registry) with the primary aim to improve care of patients with different neurological disorders including multiple sclerosis, Parkinson's disease, severe neurovascular headache, myasthenia gravis, narcolepsy, epilepsy, inflammatory polyneuropathy, as well as amyotrophic lateral sclerosis in Sweden [177, 178]. The Swedish Stroke Register is one of the world's largest stroke registers, which was established in 1994 and has included data from almost all hospitals that admit acute stroke patients in Sweden [179].

In Sweden, individual-level data in public registers are strictly protected by several laws, including the Ethics Review Act, the General Data Protection Regulation (GDPR), and the Public Access to Information and Secrecy Act (OSL). The Swedish Ethical Review Authority (Etikprövningsmyndigheten in Swedish) assesses projects according to the Ethics Review Act and requires a Swedish responsible person (Forskningshuvudman in Swedish) for the research. In addition to ethical approval, the Statistics Sweden (SCB) and the National Board of Health and Welfare (Socialstyrelsen in Swedish) also need to make an assessment according to GDPR and OSL, to determine whether individual-level data can be made available for potential research purposes. It generally takes around 1–6 months from contact person assignment to delivery of microdata in the SCB (www.scb.se/en/services/ordering-data-and-statistics/ordering-microdata/) and around 3–6 months to process applications for individual-level data in the Socialstyrelsen (www.socialstyrelsen.se/en/statistics-and-data/statistics/). According to standard legal provisions and procedures, the SCB and Socialstyrelsen only provide data to researchers working in Sweden, and researchers in other countries need to cooperate with Swedish colleagues to apply for the data.

According to the General Data Protection Regulation (GDPR), online access (e.g., through virtual machines) or transfer of individual-level data is allowed in countries of the European Union (EU) or European Economic Area (EEA), after proper legal agreements. Online access or transfer of individual-level data to an external partner in a third country outside EU/EEA is also permitted, if the third country has been approved by the European Commission and the external partner signs and complies with legal agreements that include requirements for how data must be protected, including Data Transfer Agreement (DTA), Data Processing Agreement (DPA), Material Transfer Agreement (MTA), as well as Research Collaboration Agreements.

8 Smartphone and Sensors

Smartphones and sensors allow for the unobtrusive collection of behavioral and physiological data. For instance, smartphones are commonly used in ecological momentary assessment (EMA) studies [180], resulting in continuous, real-time assessment of participant behavior, symptoms, and experiences. In addition, the built-in microphone and touchscreen of smartphones/tablets can record speech and motor movement. Recent advances in smartwatch technology has enabled many commercial devices (e.g., Fitbit, Garmin, Apple) to track physiological metrics (e.g., heart rate variability, pulse oximetry, temperature) in addition to traditional physical activity data (e.g., step count, Global Positioning System, exercise tracking). Sensors are also commonly used to collect data without requiring participant interaction. Wearable sensor devices (e.g., wrist-worn accelerometers) can collect data on sleep, activity, and physiology without burdening participants or influencing their behavior. Datasets derived from smartphone and sensor studies are typically text-based, though raw data may be proprietary. The analysis of smartphone and sensor data typically requires complex algorithms/machine learning approaches due to the complexity of data collected (in the frequency of hundreds of observations per second, from many different sensors collecting data simultaneously). Raw data is typically stored locally by the data owner, with de-identified data available upon request. In more extensive studies, data is stored and distributed through online repositories.

Several studies have collected real-world behavioral and physiological data using smartphone and sensor devices (*see* Table 2), including community twin studies (BATS, QTAB), large-scale biomedical databases (UK Biobank), and studies focusing on specific disorders (mPower).

The Brisbane Adolescent Twin Study (BATS) and the Queensland Twin Adolescent Brain (QTAB) projects are twin studies sourced from the Queensland Twin Registry (QTwin). The BATS project, enabled through funding from the NHMRC, was a longitudinal study of adolescent twins, which collected accelerometry data over three waves between 2014 and 2018 (ages 12, 14, and 16 years). The Queensland Twin Adolescent Brain study (QTAB, 2015–present), previously discussed in Subheading 5.1, collected accelerometry data over two waves (age 9–14 years at baseline). In both studies, participants wore a wrist-mounted accelerometry recording device for 2 weeks (day and night, removed only for bathing) and completed a daily sleep diary. Raw accelerometry data were processed and consolidated with sleep diary data to produce sleep onset, wake, and sleep duration estimates. The BATS and QTAB datasets include behavioral and psychological measures (e.g., assessments of cognition and behavior, self-

reported mental health and well-being) for further investigation of accelerometry measures. BATS and QTAB data is available from the project owners upon request.

The UK Biobank, previously discussed in Subheading 5.2, collected accelerometry data in 100,000 participants between 2013 and 2016. Participants wore a wrist-mounted activity monitor to capture physical activity and sleep patterns for 7 days. Since 2018, repeat measures have been collected for a subset of participants every quarter to examine seasonal influences on measurements. Data is available in raw (measured every 5 s) and average (by day and hour) acceleration formats. The deep phenotyping of the UK Biobank has allowed for accelerometry-based measures to be examined alongside several other measures, including brain structure [181], mood disorders [182], and Alzheimer's disease [183]. UK Biobank data is available online following registration (<https://bbams.ndph.ox.ac.uk/ams/>).

The mPower study (2015–present), sponsored by Sage Bionetworks with funding from the Robert Wood Johnson Foundation, aims to establish the baseline variability of real-world activity measurements of individuals with Parkinson's disease. Data is collected through an iPhone application, with minimal interruption to the daily life of participants. The initial data release (collected over 6 months) included health survey and sensor-based activity (e.g., gait and balance) data for ~8000 participants (with ~1000 self-identified as having a professional diagnosis of Parkinson's disease). In addition, approximately 900 participants contributed at least five separate days' worth of data. mPower data is accessible through the data sharing service Synapse (<https://www.synapse.org/mpower>).

A recent review [184] provides an overview of studies using smartphones to monitor symptoms of Parkinson's disease and in-depth descriptions of the methodology involved in these types of studies. Additionally, studies have used smartphone-based EMA to detect or treat mood disorders (see [185] for a review). Further, the Mobile Motor Activity Research Consortium for Health (MMARCH; <http://mmarch.org/>) is a collaborative international network working to standardize the analysis of actigraphy data in studies investigating motor activity, mood, and related disorders.

Machine learning approaches have been widely applied to data collected from smartphone and sensor devices, most notably in studies of Parkinson's disease. For example [186], used machine learning classifiers applied to accelerometry data from the UK Biobank to classify individuals with Parkinson's disease with an area under the curve of 0.85 (based on gait and low movement data). Another study [187] used data from the mPower study to detect dopaminergic medication response by applying machine learning techniques to the tapping task performance (measured via the mPower smartphone application) of Parkinson's disease patients before and after medication. Further, classifiers have been

used to detect states of deep brain stimulation (i.e., distinguishing between “On” and “Off” settings) in Parkinson’s disease patients using accelerometer and gyroscope signals from smartphones [188]. Machine learning approaches have also shown promise for other disorders. For instance, machine learning algorithms within a smartphone application have helped identify individuals with obstructive sleep apnea, using actigraphy, body position assessment, and audio recordings [189]. Lastly, some developed a pipeline for personalized modeling of depressed mood (based on EMA) and smartwatch-derived sleep and physical activity measures [190].

Acknowledgments and Fundings

This research was supported by the Australian National Health and Medical Research Council (1,078,037, 1,078,901, 1,113,400, 1,161,356, and 1,107,258), the Australian Research Council (FT180100186 and FL180100072), the Sylvia and Charles Viertel Charitable Foundation, the program “Investissements d’avenir” ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6) and reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), the European Union H2020 program (project EuroPOND, grant number 666992), the joint NSF/NIH/ANR program “Collaborative Research in Computational Neuroscience” (project HIPLAY7, grant number ANR-16-NEUC-0001-01), the ICM Big Brain Theory Program (project DYNAMO, project PredictICD), and the Abeona Foundation (project Brain@Scale). BCD is supported by a CJ Martin Fellowship (APP1161356).

References

1. UNESCO Recommendation on Open Science - UNESCO Digital Library. <https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en>
2. Fry A, Littlejohns TJ, Sudlow C et al (2017) Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol* 186:1026–1034
3. Ben-Eghan C, Sun R, Hleap JS et al (2020) Don’t ignore genetic data from minority populations. *Nature* 585:184–186
4. Yang H-C, Chen C-W, Lin Y-T et al (2021) Genetic ancestry plays a central role in population pharmacogenomics. *Commun Biol* 4: 1–14
5. Barton NTL Paul Resnick, and Genie (2019) Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. <https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>
6. Courtland R (2018) Bias detectives: the researchers striving to make algorithms fair. *Nature* 558:357–360
7. Bustamante CD, De La Vega FM, Burchard EG (2011) Genomics for the world. *Nature* 475:163–165
8. Sirugo G, Williams SM, Tishkoff SA (2019) The missing diversity in human genetic studies. *Cell* 177:26–31

9. Herrick R, Horton W, Olsen T et al (2016) XNAT central: open sourcing imaging research data. *NeuroImage* 124:1093–1096
10. Routier A, Burgos N, Díaz M et al (2021) Clinica: an open source software platform for reproducible clinical neuroscience studies. *Front Neuroinform* 15:689675
11. Petersen RC, Aisen PS, Beckett LA et al (2010) Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology* 74:201–209
12. Ellis KA, Bush AI, Darby D et al (2009) The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int Psychogeriatr* 21:672–687
13. LaMontagne PJ, Keefe S, Lauren W et al (2018) OASIS-3: longitudinal neuroimaging, clinical and cognitive dataset for normal aging and Alzheimer's disease. *Alzheimers Dement J Alzheimers Assoc* 14:P1097
14. Miller KL, Alfaro-Almagro F, Bangerter NK et al (2016) Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* 19:1523–1536
15. Alfaro-Almagro F, Jenkinson M, Bangerter NK et al (2018) Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage* 166:400–424
16. Casey BJ, Cannonier T, Conley MI et al (2018) The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. *Dev Cogn Neurosci* 32:43–54
17. Barch DM, Albaugh MD, Aveneroli S et al (2018) Demographic, physical and mental health assessments in the adolescent brain and cognitive development study: rationale and description. *Dev Cogn Neurosci* 32:55–66
18. Schmaal L, Hibar DP, Sämann PG et al (2017) Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. *Mol Psychiatry* 22:900–909
19. Hoogman M, Muetzel R, Guimaraes JP et al (2019) Brain imaging of the cortex in ADHD: a coordinated analysis of large-scale clinical and population-based samples. *Am J Psychiatry* 176:531–542
20. van Rooij D, Anagnostou E, Arango C et al (2018) Cortical and subcortical brain morphometry differences between patients with autism spectrum disorder and healthy individuals across the lifespan: results from the ENIGMA ASD Working Group. *Am J Psychiatry* 175:359–369
21. Logue MW, van Rooij SJH, Dennis EL et al (2018) Smaller hippocampal volume in post-traumatic stress disorder: a multisite ENIGMA-PGC study: subcortical volumetry results from posttraumatic stress disorder consortia. *Biol Psychiatry* 83:244–253
22. Boedhoe PSW, Schmaal L, Abe Y et al (2018) Cortical abnormalities associated with pediatric and adult obsessive-compulsive disorder: findings from the ENIGMA Obsessive-Compulsive Disorder Working Group. *Am J Psychiatry* 175:453–462
23. Mackey S, Allgaier N, Chaarani B et al (2019) Mega-analysis of gray matter volume in substance dependence: general and substance-specific regional effects. *Am J Psychiatry* 176:119–128
24. van Erp TGM, Walton E, Hibar DP et al (2018) Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the Enhancing Neuroimaging Genetics Through Meta-analysis (ENIGMA) Consortium. *Biol Psychiatry* 84:644–654
25. Hibar DP, Westlye LT, Doan NT et al (2018) Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. *Mol Psychiatry* 23:932–942
26. Dima D, Modabbernia A, Papachristou E, et al (2021) Subcortical volumes across the lifespan: data from 18,605 healthy individuals aged 3–90 years. *Hum Brain Mapp*
27. Thompson PM, Jahanshad N, Ching CRK et al (2020) ENIGMA and global neuroscience: a decade of large-scale studies of the brain in health and disease across more than 40 countries. *Transl Psychiatry* 10:1–28
28. Kremen WS, Franz CE, Lyons MJ (2013) VETSA: the Vietnam era twin study of aging. *Twin Res Hum Genet* 16:399–402
29. Van Essen DC, Smith SM, Barch DM et al (2013) The WU-Minn Human Connectome Project: an overview. *NeuroImage* 80:62–79
30. Marek K, Chowdhury S, Siderowf A et al (2018) The Parkinson's progression markers initiative (PPMI) – establishing a PD biomarker cohort. *Ann Clin Transl Neurol* 5: 1460–1477
31. Dufouil C, Dubois B, Vellas B et al (2017) Cognitive and imaging markers in non-demented subjects attending a memory clinic: study design and baseline findings of the MEMENTO cohort. *Alzheimers Res Ther* 9:67

32. Ritchie CW, Muniz-Terrera G, Kivipelto M et al (2020) The European Prevention of Alzheimer's Dementia (EPAD) Longitudinal Cohort Study: Baseline Data Release V500.0. *J Prev Alzheimers Dis* 7:8–13
33. Di Martino A, Yan C-G, Li Q et al (2014) The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol Psychiatry* 19: 659–667
34. Di Martino A, O'Connor D, Chen B et al (2017) Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci Data* 4:170010
35. Sharp, PF, Welch, A (2005) Positron emission tomography. In: Sharp PF, Gemmell HG, Murray AD (eds) Practical Nuclear Medicine. Springer, London. https://doi.org/10.1007/1-84628-018-4_3
36. Herholz K (1995) FDG PET and differential diagnosis of dementia. *Alzheimer Dis Assoc Disord* 9:6–16
37. Sudre CH, Cardoso MJ, Modat M et al (2020) Chapter 15 - Imaging biomarkers in Alzheimer's disease. In: Zhou SK, Rueckert D, Fichtinger G (eds) Handbook of medical image computing and computer assisted intervention. Academic Press, pp 343–378
38. Knudsen GM, Jensen PS, Erritzoe D et al (2016) The Center for Integrated Molecular Brain Imaging (Cimbi) database. *NeuroImage* 124:1213–1219
39. Wang H, Tian Y, Liu Y et al (2021) Population-specific brain [18F]-FDG PET templates of Chinese subjects for statistical parametric mapping. *Sci Data* 8:305
40. Jackson AF, Bolger DJ (2014) The neurophysiological bases of EEG and EEG measurement: a review for the rest of us. *Psychophysiology* 51:1061–1071
41. Niedermeyer E and da Silva FHL (2005) Electroencephalography: basic principles, clinical applications, and related fields, Lippincott Williams & Wilkins
42. Nunez PL, Srinivasan R (2006) Electric fields of the brain: the neurophysics of EEG. Oxford University Press, New York
43. Engel AK, Moll CKE, Fried I et al (2005) Invasive recordings from the human brain: clinical insights and beyond. *Nat Rev Neurosci* 6:35–47
44. Cohen D (1972) Magnetoencephalography: detection of the brain's electrical activity with a superconducting magnetometer. *Science* 175:664–666
45. Hämäläinen M, Hari R, Ilmoniemi RJ et al (1993) Magnetoencephalography---theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev Mod Phys* 65:413–497
46. Lopes da Silva F (2013) EEG and MEG: relevance to neuroscience. *Neuron* 80:1112–1128
47. Malmivuo J (2012) Comparison of the properties of EEG and MEG in detecting the electric activity of the brain. *Brain Topogr* 25:1–19
48. Bertrand O, Perrin F, Pernier J (1985) A theoretical justification of the average reference in topographic evoked potential studies. *Electroencephalogr Clin Neurophysiol* 62: 462–464
49. de Cheveigné A, Nelken I (2019) Filters: when, why, and how (not) to use them. *Neuron* 102:280–293
50. Michel CM, Brunet D (2019) EEG source imaging: a practical review of the analysis steps. *Front Neurol* 10
51. Subha DP, Joseph PK, Acharya UR et al (2010) EEG signal analysis: a survey. *J Med Syst* 34:195–212
52. Ang KK, Chin ZY, Zhang H, et al (2008) Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 2390–2397
53. Müller-Gerking J, Pfurtscheller G, Flyvbjerg H (1999) Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clin Neurophysiol* 110:787–798
54. Pfurtscheller G, Lopes da Silva FH (1999) Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin Neurophysiol* 110:1842–1857
55. Sur S, Sinha VK (2009) Event-related potential: an overview. *Ind Psychiatry J* 18:70–73
56. Neuper C, Wörz M, Pfurtscheller G (2006) ERD/ERS patterns reflecting sensorimotor activation and deactivation. *Prog Brain Res* 159:211–222
57. Crone NE, Miglioretti DL, Gordon B et al (1998) Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. *Brain J Neurol* 121(Pt 12):2301–2315
58. Bastos AM, Schoffelen J-M (2016) A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front Syst Neurosci* 9:175

59. Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci* 10:186–198
60. De Vico Fallani F, Richiardi J, Chavez M et al (2014) Graph analysis of functional brain networks: practical issues in translational neuroscience. *Philos Trans R Soc Lond Ser B Biol Sci* 369:20130521
61. Gonzalez-Astudillo J, Cattai T, Bassignana G et al (2021) Network-based brain–computer interfaces: principles and applications. *J Neural Eng* 18:0111001
62. Mirowski P, Madhavan D, LeCun Y et al (2009) Classification of patterns of EEG synchronization for seizure prediction. *Clin Neurophysiol* 120:1927–1940
63. Siddiqui MK, Morales-Menendez R, Huang X et al (2020) A review of epileptic seizure detection using machine learning classifiers. *Brain Inform* 7:5
64. Smith SJM (2005) EEG in the diagnosis, classification, and management of patients with epilepsy. *J Neurol Neurosurg Psychiatry* 76: ii2–ii7
65. Dauwels J, Vialatte F, Cichocki A (2010) Diagnosis of Alzheimer’s disease from EEG signals: where are we standing? *Curr Alzheimer Res* 7:487–505
66. Vecchio F, Babiloni C, Lizio R et al (2013) Resting state cortical EEG rhythms in Alzheimer’s disease: toward EEG markers for clinical applications: a review. *Suppl Clin Neurophysiol* 62:223–236
67. Lotte F, Bougrain L, Cichocki A et al (2018) A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *J Neural Eng* 15:031005
68. Oostenveld R, Fries P, Maris E et al (2011) FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* 2011:156869
69. Tadel F, Baillet S, Mosher JC et al (2011) Brainstorm: a user-friendly application for MEG/EEG analysis. *Comput Intell Neurosci* 2011:879716
70. Gramfort A, Luessi M, Larson E et al (2013) MEG and EEG data analysis with MNE-python. *Front Neurosci* 7:267
71. Jayaram V, Barachant A (2018) MOABB: trustworthy algorithm benchmarking for BCIs. *J Neural Eng* 15:066011
72. Carlin JB, Gurrin LC, Sterne JA et al (2005) Regression models for twin studies: a critical review. *Int J Epidemiol* 34:1089–1099
73. Sainani K (2010) The importance of accounting for correlated observations. *PM&R* 2: 858–861
74. Zagai U, Lichtenstein P, Pedersen NL et al (2019) The Swedish twin registry: content and management as a research infrastructure. *Twin Res Hum Genet* 22:672–680
75. Magnusson PKE, Almqvist C, Rahman I et al (2013) The Swedish Twin Registry: establishment of a biobank and other recent developments. *Twin Res Hum Genet* 16:317–329
76. Tomata Y, Li X, Karlsson IK et al (2020) Joint impact of common risk factors on incident dementia: a cohort study of the Swedish Twin Registry. *J Intern Med* 288:234–247
77. Wirdefeldt K, Gatz M, Pawitan Y et al (2005) Risk and protective factors for Parkinson’s disease: a study in Swedish twins. *Ann Neurol* 57:27–33
78. Fang F, Kamel F, Lichtenstein P et al (2009) Familial aggregation of amyotrophic lateral sclerosis. *Ann Neurol* 66:94–99
79. Wright MJ, Martin NG (2004) Brisbane Adolescent Twin Study: outline of study methods and research projects. *Aust J Psychol* 56:65–78
80. Miranda-Dominguez O, Feczko E, Grayson DS et al (2018) Heritability of the human connectome: a connectotyping study. *Netw Neurosci* 2:175–199
81. Kochunov P, Jahanshad N, Marcus D et al (2015) Heritability of fractional anisotropy in human white matter: a comparison of Human Connectome Project and ENIGMA-DTI data. *NeuroImage* 111:300–311
82. Schmitt JE, Raznahan A, Liu S et al (2020) The genetics of cortical myelination in young adults and its relationships to cerebral surface area, cortical thickness, and intelligence: a magnetic resonance imaging study of twins and families. *NeuroImage* 206:116319
83. Kremen WS, Franz CE, Lyons MJ (2019) Current status of the Vietnam Era Twin Study of Aging (VETSA). *Twin Res Hum Genet* 22:783–787
84. Sachdev PS, Lammel A, Trollor JN et al (2009) A comprehensive neuropsychiatric study of elderly twins: the Older Australian Twins Study. *Twin Res Hum Genet* 12:573–582
85. Hur Y-M, Bogl LH, Ordoñana JR et al (2019) Twin family registries worldwide: an important resource for scientific research. *Twin Res Hum Genet* 22:427–437
86. Ligthart L, van Beijsterveldt CEM, Kevenaar ST et al (2019) The Netherlands twin register: longitudinal research based on twin and twin-

- family designs. *Twin Res Hum Genet* 22: 623–636
87. Glahn DC, Winkler AM, Kochunov P et al (2010) Genetic control over the resting brain. *Proc Natl Acad Sci* 107:1223–1228
88. Raffield LM, Cox AJ, Hugenschmidt CE et al (2015) Heritability and genetic association analysis of neuroimaging measures in the Diabetes Heart Study. *Neurobiol Aging* 36:1602.e7–1602.15
89. Iacono WG, Heath AC, Hewitt JK et al (2018) The utility of twins in developmental cognitive neuroscience research: how twins strengthen the ABCD research design. *Dev Cogn Neurosci* 32:30–42
90. Brouwer RM, Schutte J, Janssen R et al (2021) The speed of development of adolescent brain age depends on sex and is genetically determined. *Cereb Cortex* 31:1296–1306
91. Cole JH, Poudel RPK, Tsagkrasoulis D et al (2017) Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage* 163:115–124
92. Vandenbosch MMLJZ, van't Ent D, Boomsma DI et al (2019) EEG-based age-prediction models as stable and heritable indicators of brain maturational level in children and adolescents. *Hum Brain Mapp* 40:1919–1926
93. Chung MK, Lee H, DiChristofano A et al (2019) Exact topological inference of the resting-state brain networks in twins. *Netw Neurosci* 3:674–694
94. Yamin MA, Dayan M, Squarcina L, et al (2019) Investigating the impact of genetic background on brain dynamic functional connectivity through machine learning: a Twins Study. In: 2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), pp. 1–4
95. Han Y, Adolphs R (2020) Estimating the heritability of psychological measures in the Human Connectome Project dataset. *PLoS One* 15:e0235860
96. Demeter DV, Engelhardt LE, Mallett R et al (2020) Functional connectivity fingerprints at rest are similar across youths and adults and vary with genetic similarity. *iScience* 23: 100801
97. Elliott LT, Sharp K, Alfaro-Almagro F et al (2018) Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* 562:210–216
98. Pirruccello JP, Bick A, Wang M et al (2020) Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. *Nat Commun* 11:2254
99. Liu Y, Basty N, Whitcher B et al (2021) Genetic architecture of 11 organ traits derived from abdominal MRI using deep learning. *elife* 10:e65554
100. Chua SYL, Dhillon B, Aslam T et al (2019) Associations with photoreceptor thickness measures in the UK Biobank. *Sci Rep* 9: 19440
101. Bycroft C, Freeman C, Petkova D et al (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562:203–209
102. Wain LV, Shrine N, Miller S et al (2015) Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* 3:769–781
103. Van Hout CV, Tachmazidou I, Backman JD et al (2020) Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* 586:749–756
104. Backman JD, Li AH, Marcketta A et al (2021) Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* 599:628–634
105. Wang Y, Guo J, Ni G et al (2020) Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat Commun* 11:3865
106. Medland SE, Grasby KL, Jahanshad N, et al (2020) Ten years of enhancing neuroimaging genetics through meta-analysis: an overview from the ENIGMA Genetics Working Group. *Hum Brain Mapp*
107. Adams HHH, Hibar DP, Chouraki V et al (2016) Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nat Neurosci* 19: 1569–1582
108. Hibar DP, Adams HHH, Jahanshad N et al (2017) Novel genetic loci associated with hippocampal volume. *Nat Commun* 8:13624
109. Stein JL, Medland SE, Vasquez AA et al (2012) Identification of common variants associated with human hippocampal and intracranial volumes. *Nat Genet* 44:552–561
110. Hibar DP, Stein JL, Renteria ME et al (2015) Common genetic variants influence human subcortical brain structures. *Nature* 520: 224–229
111. Satizabal CL, Adams HHH, Hibar DP et al (2019) Genetic architecture of subcortical brain structures in 38,851 individuals. *Nat Genet* 51:1624–1636

112. Grasby KL, Jahanshad N, Painter JN et al (2020) The genetic architecture of the human cerebral cortex. *Science* 367:eaay6690
113. Smit DJA, Wright MJ, Meyers JL et al (2018) Genome-wide association analysis links multiple psychiatric liability genes to oscillatory brain activity. *Hum Brain Mapp* 39:4183–4195
114. Sønderby IE, Ching CRK, Thomopoulos SI, et al (2021) Effects of copy number variations on brain structure and risk for psychiatric illness: large-scale studies from the ENIGMA working groups on CNVs. *Hum Brain Mapp*
115. Peterson RE, Kuchenbaecker K, Walters RK et al (2019) Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* 179:589–603
116. Lam M, Awasthi S, Watson HJ et al (2020) RICOPILI: rapid imputation for COsortias PIpeLIne. *Bioinforma* 36:930–933
117. Sullivan PF, Kendler KS (2021) The state of the science in psychiatric genomics. *Psychol Med* 51:2145–2147
118. Stilp AM, Emery LS, Broome JG et al (2021) A system for phenotype harmonization in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) program. *Am J Epidemiol* 190: 1977–1992
119. Talini D, Harris DN, Kessler MD et al (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature* 590:290–299
120. Patel RA, Musharoff SA, Spence JP, et al (2021) Effect sizes of causal variants for gene expression and complex traits differ between populations
121. Li Z, Li X, Zhou H, et al (2021) A framework for detecting noncoding rare variant associations of large-scale whole-genome sequencing studies
122. Schubert R, Geoffroy E, Gregg I, et al (2021) Protein prediction for trait mapping in diverse populations
123. Hindy G, Dornbos P, Chaffin MD, et al (2021) Rare coding variants in 35 genes associate with circulating lipid levels – a multi-ancestry analysis of 170,000 exomes
124. Selvaraj MS, Li X, Li Z, et al (2021) Whole genome sequence analysis of blood lipid levels in >66,000 individuals
125. Bibikova M, Barnes B, Tsan C et al (2011) High density DNA methylation array with single CpG site resolution. *Genomics* 98: 288–295
126. Yong W-S, Hsu F-M, Chen P-Y (2016) Profiling genome-wide DNA methylation. *Epigenetics Chromatin* 9:26
127. Ziller MJ, Gu H, Müller F et al (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature* 500:477–481
128. Maden SK, Thompson RF, Hansen KD et al (2021) Human methylome variation across Infinium 450K data on the gene expression omnibus. *NAR Genom Bioinf* 3:lqab025
129. Moran S, Arribas C, Esteller M (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* 8:389–399
130. Sala C, Di Lena P, Fernandes Durso D et al (2020) Evaluation of pre-processing on the meta-analysis of DNA methylation data from the Illumina HumanMethylation450 BeadChip platform. *PLoS One* 15:e0229763
131. Barrett T, Wilhite SE, Ledoux P et al (2013) NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* 41: D991–D995
132. Wang T, Guan W, Lin J et al (2015) A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. *Epigenetics* 10:662–669
133. Min JL, Hemani G, Davey Smith G et al (2018) Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinforma* 34:3983–3989
134. Birney E, Smith GD, Greally JM (2016) Epigenome-wide association studies and the Interpretation of Disease -Omics. *PLoS Genet* 12:e1006105
135. Michels KB, Binder AM (2018) Considerations for design and analysis of DNA methylation studies. *Methods Mol Biol* 1708:31–46
136. Mill J, Heijmans BT (2013) From promises to practical strategies in epigenetic epidemiology. *Nat Rev Genet* 14:585–594
137. Reed ZE, Suderman MJ, Relton CL et al (2020) The association of DNA methylation with body mass index: distinguishing between predictors and biomarkers. *Clin Epigenetics* 12:50
138. Mendelson MM, Marioni RE, Joehanes R et al (2017) Association of Body Mass Index with DNA methylation and gene expression in blood cells and relations to cardiometabolic disease: a Mendelian Randomization Approach. *PLoS Med* 14:e1002215
139. Min JL, Hemani G, Hannon E et al (2021) Genomic and phenotypic insights from an

- atlas of genetic effects on DNA methylation. *Nat Genet* 53:1311–1321
140. Chadwick LH (2012) The NIH Roadmap Epigenomics Program data resource. *Epigenomics* 4:317–324
141. Fernández JM, de la Torre V, Richardson D et al (2016) The BLUEPRINT data analysis portal. *Cell Syst* 3:491–495.e5
142. Martens JHA, Stunnenberg HG (2013) BLUEPRINT: mapping human blood cell epigenomes. *Haematologica* 98:1487–1489
143. Tryka KA, Hao L, Sturcke A et al (2014) NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Res* 42:D975–D979
144. Lonsdale J, Thomas J, Salvatore M et al (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45:580–585
145. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S et al (2020) The impact of sex on gene expression across human tissues. *Science* 369:eaba3066
146. Ferraro NM, Strober BJ, Einson J et al (2020) Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* 369:eaaz5900
147. Kim-Hellmuth S, Aguet F, Oliva M et al (2020) Cell type-specific genetic regulation of gene expression across human tissues. *Science* 369:eaaz8528
148. Demanelis K, Jasmine F, Chen LS et al (2020) Determinants of telomere length across human tissues. *Science* 369:eaaz6876
149. Burgos N, Colliot O (2020) Machine learning for classification and prediction of brain diseases: recent advances and upcoming challenges. *Curr Opin Neurol* 33:439–450
150. Koikkalainen J, Rhodius-Meester H, Tolonen A et al (2016) Differential diagnosis of neurodegenerative diseases using structural MRI data. *NeuroImage Clin* 11:435–449
151. Morin A, Samper-Gonzalez J, Bertrand A et al (2020) Accuracy of MRI classification algorithms in a tertiary memory center clinical routine cohort. *J Alzheimers Dis* 74:1157–1166
152. Rathore S, Habes M, Iftikhar MA et al (2017) A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* 155:530–548
153. Samper-González J, Burgos N, Bottani S et al (2018) Reproducible evaluation of classification methods in Alzheimer's disease: framework and application to MRI and PET data. *NeuroImage* 183:504–521
154. Wen J, Thibeau-Sutre E, Diaz-Melo M et al (2020) Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. *Med Image Anal* 63:101694
155. Daniel C, Salamanca E (2020) Hospital Databases. In: Nordlinger B, Villani C, Rus D (eds) Healthcare and artificial intelligence. Springer International Publishing, Cham, pp 57–67
156. Bottani S, Burgos N, Maire A et al (2022) Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse. *Med Image Anal* 75:102219
157. Ludvigsson JF, Otterblad-Olausson P, Pettersson BU et al (2009) The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. *Eur J Epidemiol* 24:659–667
158. Ludvigsson JF, Almqvist C, Bonamy A-KE et al (2016) Registers of the Swedish total population and their use in medical research. *Eur J Epidemiol* 31:125–136
159. Ludvigsson JF, Svedberg P, Olén O et al (2019) The longitudinal integrated database for health insurance and labour market studies (LISA) and its use in medical research. *Eur J Epidemiol* 34:423–437
160. Ekbom A (2011) The Swedish multi-generation register. *Methods Mol Biol* 675: 215–220
161. Ludvigsson JF, Andersson E, Ekbom A et al (2011) External review and validation of the Swedish national inpatient register. *BMC Public Health* 11:450
162. Song H, Sieurin J, Wirdefeldt K et al (2020) Association of stress-related disorders with subsequent neurodegenerative diseases. *JAMA Neurol* 77:700–709
163. Fang F, Zhan Y, Hammar N et al (2019) Lipids, apolipoproteins, and the risk of Parkinson Disease. *Circ Res* 125:643–652
164. Longinetti E, Mariosa D, Larsson H et al (2017) Neurodegenerative and psychiatric diseases among families with amyotrophic lateral sclerosis. *Neurology* 89:578–585
165. Barlow L, Westergren K, Holmberg L et al (2009) The completeness of the Swedish Cancer Register: a sample survey for year 1998. *Acta Oncol* 48:27–33
166. Tettamanti G, Ljung R, Ahlbom A et al (2019) Central nervous system tumor registration in the Swedish Cancer Register and Inpatient Register between 1990 and 2014. *Clin Epidemiol* 11:81–92

167. Källén B, Källén K (2003) The Swedish Medical Birth Register - a summary of content and quality. 2003:112-3
168. Persson M, Razaz N, Tedroff K et al (2018) Five and 10 minute Apgar scores and risks of cerebral palsy and epilepsy: population based cohort study in Sweden. *BMJ* 360:k207
169. Tettamanti G, Ljung R, Mathiesen T et al (2016) Maternal smoking during pregnancy and the risk of childhood brain tumors: results from a Swedish cohort study. *Cancer Epidemiol* 40:67-72
170. Brooke HL, Talbäck M, Hörnblad J et al (2017) The Swedish cause of death register. *Eur J Epidemiol* 32:765-773
171. Subic A, Zupanic E, von Euler M et al (2018) Stroke as a cause of death in death certificates of patients with dementia: a Cohort Study from the Swedish Dementia Registry. *Curr Alzheimer Res* 15:1322-1330
172. Wallerstedt SM, Wettermark B, Hoffmann M (2016) The first decade with the Swedish prescribed drug register - a systematic review of the output in the scientific literature. *Basic Clin Pharmacol Toxicol* 119:464-469
173. Wettermark B, Hammar N, Fored CM et al (2007) The new Swedish Prescribed Drug Register--opportunities for pharmacoepidemiological research and experience from the first six months. *Pharmacoepidemiol Drug Saf* 16:726-735
174. Cermakova P, Nelson M, Seznik J et al (2017) Living alone with Alzheimer's disease: data from SveDem, the Swedish Dementia Registry. *J Alzheimers Dis* 58:1265-1272
175. Haasum Y, Fastbom J, Johnell K (2016) Use of fall-risk inducing drugs in patients using Anti-Parkinson Drugs (APD): a Swedish Register-Based Study. *PLoS One* 11: e0161246
176. Religa D, Fereshtehnejad S-M, Cermakova P et al (2015) SveDem, the Swedish Dementia Registry - a tool for improving the quality of diagnostics, treatment and care of dementia patients in clinical practice. *PLoS One* 10: e0116538
177. Hillert J, Stawiarz L (2015) The Swedish MS registry – clinical support tool and scientific resource. *Acta Neurol Scand* 132:11-19
178. Longinetti E, Regodón Wallin A, Samuelsson K et al (2018) The Swedish motor neuron disease quality registry. *Amyotroph Lateral Scler Front Degener* 19:528-537
179. Asplund K, Hulter Åsberg K, Appelros P et al (2011) The Riks-stroke story: building a sustainable national register for quality assessment of stroke care. *Int J Stroke* 6:99-108
180. Shiffman S, Stone AA, Hufford MR (2008) Ecological momentary assessment. *Annu Rev Clin Psychol* 4:1-32
181. Hamer M, Sharma N, Batty GD (2018) Association of objectively measured physical activity with brain structure: UK Biobank study. *J Intern Med* 284:439-443
182. Lyall LM, Wyse CA, Graham N et al (2018) Association of disrupted circadian rhythmicity with mood disorders, subjective wellbeing, and cognitive function: a cross-sectional study of 91 105 participants from the UK Biobank. *Lancet Psychiatry* 5:507-514
183. Huang J, Zuber V, Matthews PM et al (2020) Sleep, major depressive disorder, and Alzheimer disease: a Mendelian randomization study. *Neurology* 95:e1963-e1970
184. Little MA (2021) Smartphones for remote symptom monitoring of Parkinson's disease. *J Parkinsons Dis* 11:S49-S53
185. Yim SJ, Lui LMW, Lee Y et al (2020) The utility of smartphone-based, ecological momentary assessment for depressive symptoms. *J Affect Disord* 274:602-609
186. Williamson JR, Telfer B, Mullany R et al (2021) Detecting Parkinson's disease from Wrist-Worn Accelerometry in the U.K. Biobank. *Sensors* 21:2047
187. Chaibub Neto E, Bot BM, Perumal T et al (2016) Personalized hypothesis tests for detecting medication response in Parkinson disease patients using iPhone sensor data. *Pac Symp Biocomput Pac Symp Biocomput* 21:273-284
188. LeMoine R, Mastroianni T, Whiting D et al (2019) Assessment of machine learning classification strategies for the differentiation of deep brain stimulation “on” and “off” status for Parkinson's disease using a smartphone as a wearable and wireless inertial sensor for quantified feedback. In: LeMoine R, Mastroianni T, Whiting D et al (eds) *Wearable and Wireless Systems for Healthcare II: movement disorder evaluation and deep brain stimulation systems*. Springer, Singapore, pp 113-126
189. Behar J, Roebuck A, Shahid M et al (2015) SleepAp: an automated obstructive sleep apnoea screening application for smartphones. *IEEE J Biomed Health Inform* 19: 325-331
190. Shah RV, Grennan G, Zafar-Khan M et al (2021) Personalized machine learning of depressed mood using wearables. *Transl Psychiatry* 11:1-18
191. Vasanthakumar A, Davis JW, Idler K et al (2020) Harnessing peripheral DNA methylation differences in the Alzheimer's Disease

- Neuroimaging Initiative (ADNI) to reveal novel biomarkers of disease. *Clin Epigenetics* 12:84
192. Cho H, Ahn M, Ahn S et al (2017) EEG datasets for motor imagery brain-computer interface. *GigaScience* 6:1–8
193. Cattan G, Rodrigues PLC, Congedo M (2018) EEG Alpha Waves Dataset. <https://hal.archives-ouvertes.fr/hal-02086581>
194. Abel JH, Badgeley MA, Meschede-Krasa B et al (2021) Machine learning of EEG spectra classifies unconsciousness during GABAergic anesthesia. *PLoS One* 16:e0246165
195. Wakeman DG, Henson RN (2015) A multi-subject, multi-modal human neuroimaging dataset. *Sci Data* 2:150001
196. Blankertz B, Dornhege G, Krauledat M et al (2007) The non-invasive Berlin Brain-Computer Interface: fast acquisition of effective performance in untrained subjects. *NeuroImage* 37:539–550
197. Tangermann M, Müller K-R, Aertsen A et al (2012) Review of the BCI Competition IV. *Front Neurosci* 6:55
198. Mumtaz W, Xia L, Yasin MAM et al (2017) A wavelet-based technique to predict treatment outcome for Major Depressive Disorder. *PLoS One* 12:e0171409
199. Mohammadi MR, Khaleghi A, Nasrabadi AM et al (2016) EEG classification of ADHD and normal children using non-linear features and neural network. *Biomed Eng Lett* 6:66–73
200. Jaramillo-Gonzalez A, Wu S, Tonin A et al (2021) A dataset of EEG and EOG from an auditory EOG-based communication system for patients in locked-in state. *Sci Data* 8:8
201. Shah V, von Weltin E, Lopez S et al (2018) The temple university hospital seizure detection corpus. *Front Neuroinformatics* 12:83
202. Andrzejak RG, Schindler K, Rummel C (2012) Nonrandomness, nonlinear dependence, and nonstationarity of electroencephalographic recordings from epilepsy patients. *Phys Rev E Stat Nonlinear Soft Matter Phys* 86:046206
203. Goldberger AL, Amaral LA, Glass L et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101:E215–E220
204. Brunner C, Birbaumer N, Blankertz B et al (2015) BNCI Horizon 2020: towards a roadmap for the BCI community. *Brain-Comput Interfases* 2:1–10
205. O'Callaghan VS, Hansell NK, Guo W et al (2021) Genetic and environmental influences on sleep-wake behaviours in adolescence, vol 2. *SLEEP Adv*, p zpab018
206. de Zubicaray GI, Chiang M-C, McMahon KL et al (2008) Meeting the challenges of neuro-imaging genetics. *Brain Imaging Behav* 2: 258–263
207. Relton CL, Gaunt T, McArdle W et al (2015) Data resource profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int J Epidemiol* 44:1181–1190
208. Bonder MJ, Luijk R, Zhernakova DV et al (2017) Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet* 49:131–138
209. Huan T, Joehanes R, Song C et al (2019) Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nat Commun* 10:4267
210. Sarnowski C, Satizabal CL, DeCarli C et al (2018) Whole genome sequence analyses of brain imaging measures in the Framingham study. *Neurology* 90:e188–e196
211. Westerman K, Sebastiani P, Jacques P et al (2019) DNA methylation modules associate with incident cardiovascular disease and cumulative risk factor exposure. *Clin Epigenetics* 11:142
212. Nabais MF, Lin T, Benyamin B et al (2020) Significant out-of-sample classification from methylation profile scoring for amyotrophic lateral sclerosis. *Npj Genomic Med* 5:1–9
213. Vallerga CL, Zhang F, Fowdar J et al (2020) Analysis of DNA methylation associates the cystine–glutamate antiporter SLC7A11 with risk of Parkinson's disease. *Nat Commun* 11: 1238
214. Sullivan PF (2010) The psychiatric GWAS consortium: big science comes to psychiatry. *Neuron* 68:182–186
215. Sullivan PF, Agrawal A, Bulik CM et al (2018) Psychiatric genomics: an update and an agenda. *Am J Psychiatry* 175:15–27
216. Mak ACY, White MJ, Eckalbar WL et al (2018) Whole-genome sequencing of pharmacogenetic drug response in racially diverse children with asthma. *Am J Respir Crit Care Med* 197:1552–1564
217. Kurniansyah N, Goodman MO, Kelly T, et al (2021) A multi-ethnic polygenic risk score is associated with hypertension prevalence and progression throughout adulthood, medRxiv
218. Hu Y, Stilp AM, McHugh CP et al (2021) Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: the NHLBI TOPMed program. *Am J Hum Genet* 108:874–893

- 219. Wu K-D, Hsiao C-F, Ho L-T et al (2002) Clustering and heritability of insulin resistance in Chinese and Japanese hypertensive families: a Stanford-Asian Pacific Program in Hypertension and Insulin Resistance sibling study. *Hypertens Res* 25:529–536
- 220. Johnsen JM, Fletcher SN, Huston H et al (2017) Novel approach to genetic analysis and results in 3000 hemophilia patients enrolled in the my life, our future initiative. *Blood Adv* 1:824–834
- 221. Zhao X, Qiao D, Yang C et al (2020) Whole genome sequence analysis of pulmonary function and COPD in 19,996 multi-ethnic participants. *Nat Commun* 11:5182
- 222. GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369:1318–1330
- 223. Sletten TL, Rajaratnam SMW, Wright MJ et al (2013) Genetic and environmental contributions to sleep-wake behavior in 12-year-old twins. *Sleep* 36:1715–1722
- 224. Mitchell BL, Campos AI, Rentería ME et al (2019) Twenty-five and up (25Up) study: a new wave of the Brisbane Longitudinal Twin Study. *Twin Res Hum Genet* 22:154–163
- 225. Zietsch BP, Hansen JL, Hansell NK et al (2007) Common and specific genetic influences on EEG power bands delta, theta, alpha, and beta. *Biol Psychol* 75:154–164
- 226. Bot BM, Suver C, Neto EC et al (2016) The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci Data* 3: 160011

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part V

Disorders



Chapter 25

Machine Learning for Alzheimer's Disease and Related Dementias

Marc Modat, David M. Cash, Liane Dos Santos Canas, Martina Bocchetta, and Sébastien Ourselin

Abstract

Dementia denotes the condition that affects people suffering from cognitive and behavioral impairments due to brain damage. Common causes of dementia include Alzheimer's disease, vascular dementia, or frontotemporal dementia, among others. The onset of these pathologies often occurs at least a decade before any clinical symptoms are perceived. Several biomarkers have been developed to gain a better insight into disease progression, both in the prodromal and the symptomatic phases. Those markers are commonly derived from genetic information, biofluid, medical images, or clinical and cognitive assessments. Information is nowadays also captured using smart devices to further understand how patients are affected. In the last two to three decades, the research community has made a great effort to capture and share for research a large amount of data from many sources. As a result, many approaches using machine learning have been proposed in the scientific literature. Those include dedicated tools for data harmonization, extraction of biomarkers that act as disease progression proxy, classification tools, or creation of focused modeling tools that mimic and help predict disease progression. To date, however, very few methods have been translated to clinical care, and many challenges still need addressing.

Key words Dementia, Alzheimer's disease, Cognitive impairment, Machine learning, Data harmonization, Biomarkers, Imaging, Classification, Disease progression modeling

1 Introduction

Dementia is a progressive condition which affects over 55 million people worldwide, with nearly 10 million new cases every year [1]. The term “dementia” indicates not a single disease, but rather a spectrum of different conditions with different clinical phenotypes, which can be caused by a multitude of pathologies that cause changes in the structure and chemistry of the brain. While the most common cause of dementia-related symptoms is a neurodegenerative disease, other causes do exist (e.g., chronic inflammatory disease, alcoholism...). The exact pathological cascade of events which causes the development of symptoms is still unknown, but overall it

is thought that a combination of genetic and environmental factors results in the abnormal accumulation of misfolded, toxic proteins in the brain, which then triggers both chemical imbalance and neuronal loss in the brain (a process called atrophy), ultimately leading to the hallmark clinical symptoms that eventually impair the daily functioning of affected individuals. An important distinction to make is between the concept of “dementia” as a collection of clinical syndromes and as qualitative and quantitative clinical expressions of the disease, and “disease” as the underlying pathophysiological processes of the syndromes.

Thanks to the increased insight into disease pathophysiology, there has been a revision of the clinical diagnostic criteria, moving from considering the observable clinical signs and symptoms and implying a close and consistent correspondence between clinical symptoms and the underlying pathology, to including biomarkers of the underlying disease state in the clinical diagnosis. For example, the 1984 NINCDS-ADRDA¹ criteria were the benchmark for a clinical diagnosis of Alzheimer’s disease, which was defined as “a progressive, dementing disorder, usually of middle or late life” [2]. These criteria were revised in 2011 [3], to include biomarkers to support the clinical diagnosis and to account for the “pre-dementia” stages and the slow pathological changes occurring over many years before the manifestation of clinical symptoms [4].

Despite different pathological origins, many forms of dementia can have similar symptoms, which typically include memory loss, language difficulties, disorientation, and behavioral changes. However, at an individual level, the symptoms can vary with regard to their nature, presentation, rate of progression, and severity. Such heterogeneity between and within forms of dementia is typically related to the area (or areas) of the brain affected by the underlying pathology and by the etiological cause of the disease itself.

1.1 Alzheimer’s Disease (AD)

AD is the most common form of dementia, accounting for 60–65% of all cases. It typically presents in individuals aged 65 or older, with the initial and most prominent cognitive deficits being memory loss, with additional cognitive impairments in the language, visuospatial, and executive functions [3]. The distinguishing feature of AD is the buildup of amyloid- β plaques and neurofibrillary tangles of tau proteins. The amyloid plaques tend to be diffuse throughout the brain, while tau pathology tends to start in the mediotemporal lobe, and in particular in the hippocampus and entorhinal cortex, and spread to prefrontal and temporoparietal cortex in the moderate stages of the disease. There are numerous genetic factors that have different levels of risk and prevalence in the population. The

¹ National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer’s Disease and Related Disorders Association.

greatest risk comes from the nearly fully penetrant autosomal dominant mutations in the amyloid precursor protein (*APP*), presenilin 1 (*PSEN1*), or presenilin 2 (*PSEN2*) genes. However, the prevalence of these mutations is extremely low, comprising less than 0.5% of all AD cases. The age at onset of autosomal dominant AD is relatively similar between generations [5] and within individual mutations [6], typically resulting in an early-onset form of AD (below the age of 60 years). The most prominent risk factor gene in terms of both hazard and prevalence is apolipoprotein E (*APOE*). Carriers of a single copy (roughly 25% of the population) of the $\epsilon 4$ allele are roughly two to three times more likely to develop AD [7], and they tend to have an earlier age of disease onset. Homozygotic $\epsilon 4$ carriers represent 2–3% of the general population, with a dose-dependent increase in risk. There have been some suggestions that carrying an $\epsilon 2$ form of *APOE* can infer some protection to individuals compared to the most common $\epsilon 3$ allele [7, 8].

Besides the typical presentations of AD in which episodic memory deficits are prominent, there are other variants with atypical presentations. Posterior cortical atrophy (PCA) [9] is characterized by visual and spatial impairments, but memory and language abilities are preserved in the early stages, with atrophy localized in the parietal and occipital lobe. Logopenic variant of the primary progressive aphasia (lvPPA), also called logopenic progressive aphasia (LPA), is characterized by impairments in the language domain (i.e., word-finding difficulty, impaired repetition of sentences and phrases) and atrophy in the left temporoparietal junction [10]. Despite presenting with different symptoms and neuroanatomical features, both PCA and lvPPA typically share the same forms of pathology, amyloid plaques, and neurofibrillary tangles, with the typical forms of AD. Besides these pathological hallmarks, accumulation of the TAR DNA-binding protein 43 (TDP-43) [11] is another form of pathology often observed in AD, particularly in cases with older onset of symptoms, resulting in increased rates of atrophy. The limbic-predominant age-related TDP-43 encephalopathy dementia (LATE) is a related condition found in older elderly adults (above 80 years of age), presenting with a slow progression of amnestic symptoms and hippocampal sclerosis.

1.2 Vascular Dementia (VaD)

As the second most common form of dementia (accounting for 10–15% of all dementia cases), VaD is an umbrella term for a number of syndromes due to a clear primary cause: the decreased blood flow due to damage in the blood supply (large or small vessels), which leads to brain tissue damage. The vascular origin is clearly seen on magnetic resonance imaging (MRI) as the presence of extensive periventricular white matter lesions, or multiple lacunes in the basal ganglia and/or white matter [12]. Symptoms tend to accumulate in a step-wise fashion, rather than gradually

worsening, and they greatly vary based on which vessel is involved: ranging from memory loss, and difficulties in executive functions, to language and motor impairments. Different syndromes include multi-infarct dementia (or vascular cognitive impairment), when a series of small strokes damage multiple areas of the brain, typically in the cortex; strategic infarct dementia, when symptoms are caused by a focal ischemic lesion; subcortical vascular dementia (or subcortical leukoencephalopathy), caused by occlusions in small vessels, resulting in multiple lacunes in the subcortical structures; and mixed dementia, when symptoms of both vascular dementia and AD are present.

Generally, strategic infarct dementia and multi-infarct dementia involving the cortex are due to occlusion in one of the major cerebral arteries, and therefore the insult in the brain usually results in a large area affected; they have a definable time of onset and specific deficits related to the region affected. When the occlusion involves small vessels, the dementia symptoms have a more insidious onset and less defined deficits in the executive function domain.

Risk factors typically include age, hypertension, high cholesterol, obesity, smoking, and other cardiovascular diseases (family history of stroke, heart disease, or diabetes). Mutations in the Notch3 gene have been associated to the cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL), which is a genetic disorder showing recurrent stroke, resulting in lacunar infarcts [13].

1.3 Frontotemporal Dementia (FTD)

FTD describes a very heterogeneous group of neurodegenerative disorders with multiple genetic and pathological causes. However, there is sufficient overlap in terms of both clinical (behavioral and/or language symptoms) and anatomical presentation (frontal and temporal lobe atrophy and hypometabolism) that the conditions are commonly considered together as one group. While representing 5–10% of all dementia cases, the FTD disorders constitute a more common cause of early onset dementia, approximately equal in frequency to AD in people under the age of 65. The only confirmed risk factors are genetic, and about a 30–50% of cases are due to an autosomal dominant mutation, primarily found in the microtubule-associated protein tau (*MAPT*), progranulin (*GRN*), or chromosome 9 open reading frame 72 (*C9orf72*) genes [14]. The age at onset is extremely variable within and between genetic forms, including within families, and therefore hard to predict [15].

Clinically, behavioral variant FTD (bvFTD) is the most common presentation, with impaired social conduct and personality changes, often misdiagnosed as psychiatric illness at the onset [16]. It could be caused by tau, TDP-43, or fused-in-sarcoma pathology [17, 18] and associated with extremely variable pattern

of atrophy between patients, with a predominant involvement of the frontal and temporal cortex (often asymmetrical), but also insula, anterior cingulate, and subcortical structures [19–21].

Less frequently, patients present with progressive decline in speech and language functions, a collection of disorders and variants referred to as primary progressive aphasia, PPA. There are multiple variants of PPA, with different language deficits and brain regions involved. These are semantic variant (svPPA), with a breakdown of semantic memory, and associated atrophy in the left antero-inferior temporal lobe [22, 23]; non-fluent variant (nfvPPA), characterized by agrammatism and speech apraxia, and associated atrophy in the left inferior frontal, superior temporal, and insular cortex [22]; and lvPPA, though as mentioned previously is more often linked with AD pathology [10].

Around 15% of people on the FTD spectrum can also develop motor features consistent with either amyotrophic lateral sclerosis (ALS) (or motor neurone disease, MND) or parkinsonism (including progressive supranuclear palsy, PSP, or corticobasal syndrome, CBS) [24].

There is a distinct differential brain involvement across the genetic forms of FTD, evident up to 15 years before the estimated symptoms onset [25, 26]: *MAPT* mutations cause focal symmetric atrophy in the anterior temporal and orbitofrontal cortex, including hippocampus and amygdala; *GRN* mutations usually cause asymmetric atrophy in the temporal, inferior frontal, and inferior parietal lobes and striatum; while *C9orf72* repeated expansions showed wider symmetric atrophy, predominantly involving the dorsolateral and medial frontal and orbitofrontal cortex, as well as the thalamus and cerebellum [25, 27, 28]. Despite these common patterns, there is still large variability even within the same genetic group, potentially due to the specific mutations, clinical presentations, or genetic and environmental factors [29, 30].

1.4 Dementia with Lewy Bodies (DLB)

Around 10–15% of dementia cases have a diagnosis of DLB [31]. Symptoms tend to have an insidious onset, usually at the age of 65 years or older, and disease duration has an average of 5 to 8 years from diagnosis, but it can range from 2 to 20 years. Symptoms change greatly from person to person but typically include fluctuating cognition, pronounced alterations in attention, alertness and executive functions, visual hallucinations, and motor features of parkinsonism. Early signs also include rapid eye movement (REM) sleep behavior disorder, while memory and hippocampal volume are relatively preserved in the initial stages, but they become impaired later during the course of the disease. Alongside relatively preserved mediotemporal lobe volumes, typical biomarkers are reduced dopamine transporter (DAT) uptake in the basal ganglia on single-photon emission computed tomography (SPECT) or positron emission tomography (PET) imaging, and polysomnographic recordings, showing REM sleep without atonia.

DLB is considered a sporadic disease; however, mutations in genes encoding α -synuclein (SNCA) and β -synuclein (SNCAβ) proteins have been associated with DLB [32].

Pathologically, DLB is characterized by the presence of α -synuclein proteins which abnormally aggregate in the brain to form Lewy bodies.

Lewy bodies are also found in the brain of individuals affected by Parkinson's disease and Parkinson's disease dementia (PDD). DLB and PDD are often difficult to distinguish, and the "1-year rule" is used for differential diagnosis: if the parkinsonian motor symptoms are experienced for a year or more before the onset of the cognitive impairments, then the condition of PDD is diagnosed, while if the cognitive problems start before or within 1 year after the movement difficulties, then a diagnosis of DLB is likely to be given.

Box 1: Different Diseases Causing Dementia

Dementia is not a disease but a spectrum of disorders defined by different pathologies, the most common being the following.

- Alzheimer's disease is the most prevalent, with hallmark pathologies of amyloid- β and neurofibrillary tau tangles. Memory is the most common symptom, but there are visual, language, and behavioral variants.
- Vascular dementia is caused by various types of vascular insults.
- Frontotemporal dementia is more common in those with younger ages of onset and is more associated with behavioral and language forms.
- Dementia with Lewy bodies (DLB) shares pathology with parkinsonian disorders and often has visual fluctuations and hallucinations as symptoms.

2 Features and Markers of Dementia

As previously mentioned, the most prevalent forms of dementia are multi-factorial processes that typically occur over a very long time period, from the silent buildup of pathology through to the onset and progression of the clinical syndrome. As such, there will be numerous types of assessment that can help identify individuals at risk, underlying pathology burden, and severity of the disease. These range from classic clinical workups, cognitive assessments of memory and other brain functions, fluid-based biomarkers, and

medical imaging-based assessments. The utility and sensitivity of these investigations will highly depend on the stage of the disease that the patient is experiencing.

2.1 Genetic Markers

Despite the numerous forms of pathology that can ultimately lead to dementia, many genetic risk factors have been identified for AD and related disorders; the overall heritability of AD has been characterized to be between 60 and 80% [33]. This risk of heritability however is spread out over a wide range of locations that vary in terms of prevalence and impact. Identifying genetic risk factors and the associated pathways that these genes are involved in have led to a better understanding of various forms of dementia [34].

As mentioned in Subheading 1, the genetic variants with the strongest penetrance are the autosomal dominant forms of dementia. What these rare autosomal dominant forms provide is an opportunity to study a “purer” form of dementia, as the age of disease onset tends to be in the 30s through 50s, when there should be a far lower likelihood of comorbidities. It also provides a chance to study pre-symptomatic changes in individuals who are nearly certain to become affected by the disease. Thus, these cohorts are an ideal population for clinical trials of new therapies, in part to prove that the target engagement is successful and whether it provides any evidence that supports the underlying hypothesis around the disease start and spread.

Outside of the autosomal dominant mutations, the gene most linked with risk for AD is *APOE*. There is not an equivalent gene in terms of risk and prevalence to *APOE* yet discovered for other forms of dementia, in part because these forms of dementia are rarer and it is thus more difficult to include the number of subjects needed for a well-powered GWAS (genome-wide association study). However, there are some suggestions, such as the *TREM2* variant in FTD [35].

Rather than trying to identify single target genes and their associated risks, many researchers have looked to generate a polygenic risk score, i.e., a sum of the risks conferred by each associated variant across the genome. Polygenic risk scores (PRS) have been developed for multiple diseases to better account for the amalgamated risk that the entire genetic profile provides [36]. For AD, however, *APOE* confers a far greater risk to individuals, with the PRS scores able to slightly improve predictive accuracy and explain additional risks beyond *APOE* [37, 38].

2.2 Clinical and Cognitive Assessment

Given that various forms of dementia have historically been defined by their clinical phenotype, and that clinical and cognitive assessments tend to be the cheapest and most widely available, they often are paramount in terms of initial diagnostic workup of an individual, as well as their subsequent patient management.

Clinical vital signs, such as blood pressure [39, 40] and body mass index (BMI) [41], may suggest causes of cognitive impairment other than a neurodegenerative disorder or indications of an at-risk profile that would result in a more aggressive disease. The Clinical Dementia Rating (CDR) [42] is a semi-structured interview that examines several aspects of physical and mental well-being which is summarized under six subdomains. For each subdomain, a score of 0 (no impairment), 0.5 (mild/questionable impairment), 1, 2, or 3 is given. Both the sum of these subdomains, referred to as the CDR Sum of Boxes (CDR-SB), and a global summary score are often used, with CDR-SB now commonly used as a primary endpoint in trials. Other clinical workups may help identify non-memory symptoms that would not be picked up via cognitive assessments, such as anxiety, depression, and quality of daily activities.

Cognitive assessments look at numerous domains of brain function, including executive function, language, visuospatial functions, and behavior. However, given that memory is the most common primary complaint from individuals with AD, assessments of various aspects of an individual's memory is one of the most important and typically included in both clinical and research settings. Numerous tests have been developed and validated for use in the clinic as well, and they often serve as a primary outcome measure in clinical trials of subjects with mild to moderate AD. Standard clinical and cognitive assessments include the Mini Mental State Exam (MMSE) [43], the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-COG) [44], and the Montreal Cognitive Assessment (MoCA) [45].

While cheap and readily available, these assessments do come with some disadvantages. These are often pencil and paper tests which are administered and scored by a trained rater. As such, there is a level of subjectivity in many of these assessments that tend to result in high variability. Often these tests repeat the same questions and tasks over and over again, which leads to practice effects. It also is often difficult to build these assessments such that their dynamic range can simultaneously cover both the early subtle signs of dementia pre-symptomatically and the full decline once the individuals have experienced symptoms. This results in some tests having substantial ceiling effects (i.e., being easy enough that there is limited distinction between healthy individuals and those experiencing the subtle initial symptoms) and floor effects (i.e., the tests are so difficult that many with a cognitive impairment cannot perform them). There are also the cultural and lingual artifacts that may produce bias when translating one of these tests over from one language to another. As a result, there is a trend to formulate cognitive assessments in a more objective, computational format to reduce issues around subjectivity, language differences, and learning effects. They may reduce the variability compared to

standard paper and pen tests, which may be of key benefit in assessing therapeutic effects in clinical trials [46, 47]. This includes multiple trials of the test run during a single assessment and collection of dense information about the task in addition to more summary metrics as number of items correct or mean reaction time. The rich set of detailed, repeated measures is ideal for further exploration with machine learning algorithms.

2.3 Biofluid Assessments

The most widely studied fluid-based biomarkers in AD and related disorders come from samples of cerebrospinal fluid extracted from an individual's spine. Measures of primary AD-related pathology ($A\beta_{1-42}$, tau, p-tau) can be obtained from these samples, as well as peripheral information on downstream mechanisms, such as neuroinflammation, synaptic dysfunction, and neuronal injury [48]. Fluid-based biomarkers are very effective in terms of being a "state" biomarker, i.e., whether an individual has a normal or abnormal level. They are often much cheaper than imaging assessments in providing this status and thus are more likely to be used for screening of individuals at risk for dementia. At the same time, their ability to track change in the disease over time is currently limited. They are in general more noisy measurements, likely due to a number of factors including consistency of extraction, storage, and analysis methods [49]. Even when these have been held extremely consistent, their variability is still much higher in terms of measurement of change over time compared to cognitive and imaging measures [50]. Despite the procedure being very safe and continuing to improve, there is still a set of individuals who will not wish to participate in studies involving these assessments. A far less invasive and cheaper procedure is to extract similar measures from the plasma. While plasma-based biomarkers have been actively pursued for a lengthy time, it is only very recently that they have produced the level of accuracy and precision needed to compete in terms of performance to other established measurements [51, 52]. There have been plasma-based assessments of amyloid- β , different tau isoforms, and nonspecific markers of neurodegeneration (such as levels of the neurofilament light chain, NfL) which show promise for detecting changes in the preclinical stage of AD [53].

2.4 Imaging Dementia

The primary use of brain imaging in clinical settings is to exclude non-neurodegenerative causes, such as normal pressure hydrocephalus, tumors, and chronic hemorrhages, together with absence of atrophy, all features that can be visualized on T1-weighted MRI or computerized tomography (CT) scans. Nevertheless, three-dimensional tomographic medical imaging modalities, particularly PET and MR imaging, provide high-precision measurements of spatiotemporal patterns of disease burden that have proven extremely valuable for research and also currently contribute to the positive diagnosis.

These modalities have been employed primarily in clinical research settings, where longer advanced imaging protocols and novel radiotracers can be implemented. Due to costs, time, and availability of imaging resources, they have been slow to translate to the clinical setting itself, but they are beginning to make impact there as well.

2.4.1 Imaging Primary Pathology

Radiotracers that preferentially bind to the primary pathologies associated with AD allow the detection and tracking of the slow progressive buildup of the amyloid plaques and neurofibrillary tangles, which can occur decades before the onset of symptoms [54, 55]. The original amyloid tracer was ¹¹C Pittsburgh Compound B [PIB], and it identified individuals who were amyloid positive but showed no symptoms [56]. These individuals tended to progress to mild cognitive impairment and subsequently AD at much higher rates than those who were amyloid negative [57]. Since the introduction of PIB, there have been numerous ¹⁸F tracers which have been developed and approved for use in humans [58, 59]. As ¹⁸F-based tracers have a longer half-life than ¹¹C, it has enabled a much larger group of research centers access to this technology. Tracers specifically related to tau-based pathology have come much later. The most widely used has been florlaucipir, with second-generation tracers now available that have overcome some of the challenges of imaging with the early tau tracers [60]. Findings from tau PET studies suggest that the landmark postmortem staging of tau pathology seeding and spread according to Braak [61] is the most common spatiotemporal pattern observed in individuals [62, 63]. However, other subtypes of different distributions have been observed [64, 65]. Elevated tau PET uptake often happens much later than elevation of amyloid PET [66], especially in autosomal dominant cases of AD [67, 68]. Tau PET is also far more strongly linked regionally with subsequent evidence of neurodegeneration, while amyloid PET tends to elevate in a similar manner across multiple regions at the same time [69]. Examples of amyloid and tau PET images from both patients with various forms of AD and controls can be seen in Fig. 1. Despite many forms of FTD being some form of tauopathy, the available PET tracers have been primarily optimized to the specific form of tau pathology that is primarily observed in AD, namely, the mix of 3-Repeat/4-Repeat species observed in neurofibrillary tangles. Since there are many different forms of tau pathology within FTD, the level of tau PET uptake in these individuals is varied [70–72]. In other forms of dementia, amyloid PET can be used to rule out AD pathology if an individual with symptoms has an amyloid negative scan² and tau

² https://www.accessdata.fda.gov/drugsatfda_docs/nda/2012/202008_Florbetapir_Orig1s000TOC.cfm.

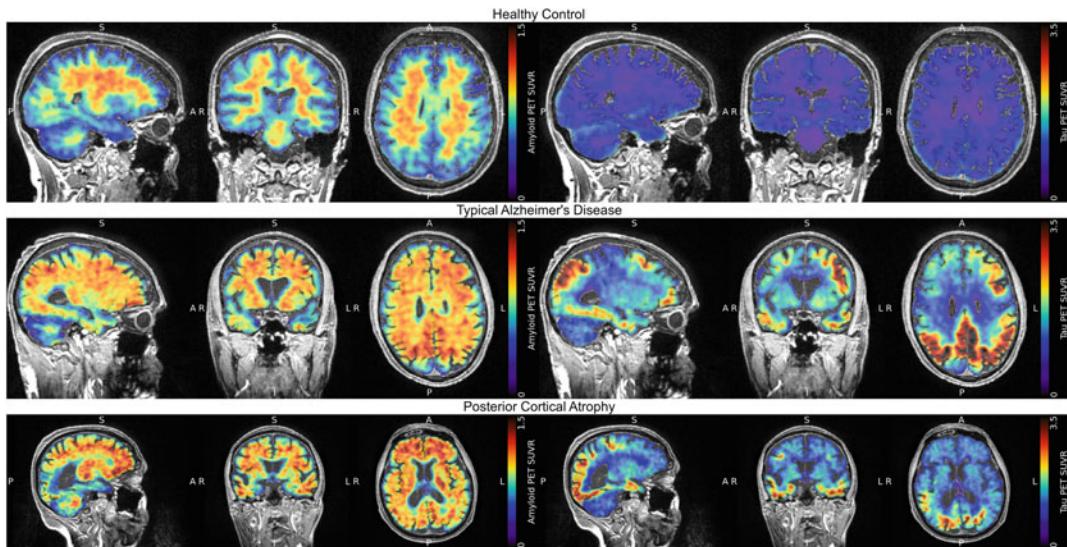


Fig. 1 Example amyloid (left column) and tau (right column) PET scans from controls and patients with different forms of dementia (AD and PCA). PET images are presented as standardized uptake value ratio (SUVR) images, where the amyloid PET have been normalized to subcortical WM, an area of high nonspecific binding (mainly in myelin) for both healthy controls and patients with Alzheimer's disease. The tau PET images have been normalized to inferior cerebellar gray matter. While amyloid PET tends to show diffuse cortical uptake across the brain, tau PET tends to be more focal in the areas where neurodegeneration is occurring

PET has now been approved to estimate the density and distribution of neurofibrillary tangles in individuals.³

2.4.2 Imaging Neurodegeneration

As the pathology continues to build over time during the pre-symptomatic period, it often leads to an insidious process of neuronal dysfunction and ultimately to degeneration in all forms of dementia. This is evidenced by atrophy visible in the structural T1-weighted MRI scans (Fig. 2) and decreased metabolism on fluorodeoxyglucose (FDG)-PET (Fig. 3). These forms of imaging start to be altered around the time when tau pathology is present and then provide close tracking with disease severity as symptoms become apparent. These modalities often tend to be the most widely available of imaging techniques within research settings, with MRI tending to be less costly than PET. Structural imaging, due to its high resolution (1 mm), signal-to-noise ratio, and contrast between tissues, lends itself to high-precision measurements of change over time. The spatial pattern of the neurodegeneration, whether it is hypometabolism or atrophy, can provide useful information for differential diagnosis between different dementia [73–75]. Parallel to neurodegeneration, changes in the white matter of individuals with dementia also show evidence of disease-related

³ https://www.accessdata.fda.gov/drugsatfda_docs/label/2020/212123s000lbl.pdf.

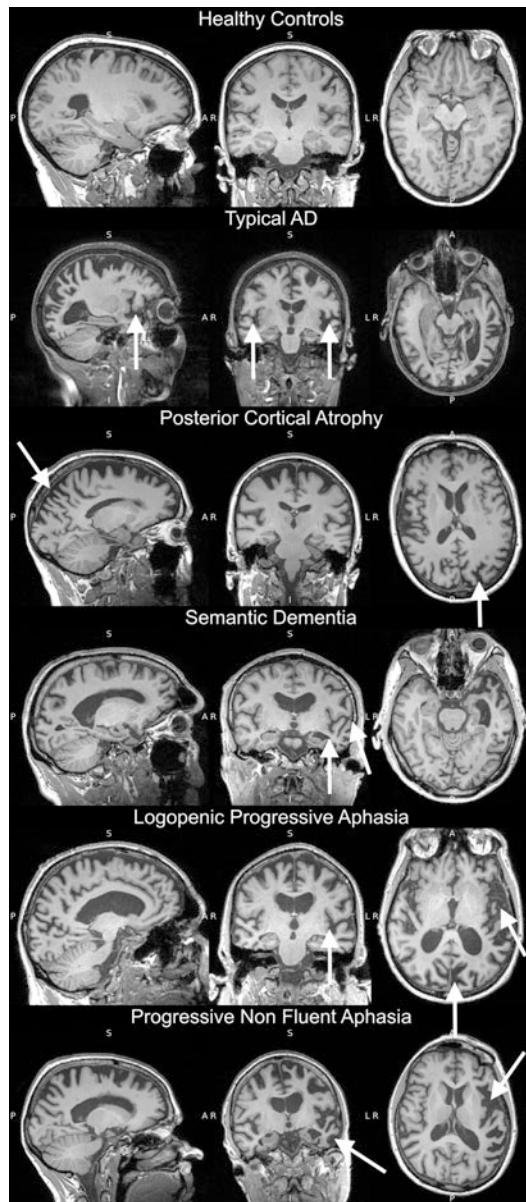


Fig. 2 Example T1-weighted MRI scans from a healthy control and individuals with different variants of Alzheimer's disease. For each variant, atrophy can be observed in areas of cortical GM which are known to cause the cognitive deficits typically linked to the clinical phenotype (see white arrows)

insult. White matter lesions, suggestive of damage due to vascular insult/insufficiency or demyelination, are visible as hypointensities on T1-weighted imaging, while they present as hyperintense on other forms of structural MRI imaging known as T2-weighted or fluid attenuated inversion recovery (FLAIR) (Fig. 4). Other forms of changes in the WM observed in dementia include microbleeds,

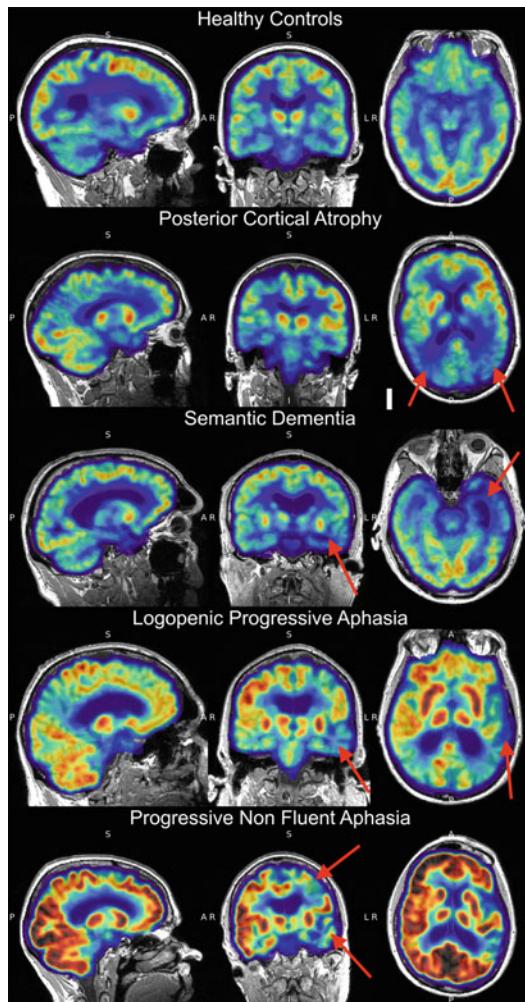


Fig. 3 Example FDG PET scans from a healthy control and multiple variants of Alzheimer's disease. For each variant, hypometabolism (denoted by cooler colors) observed in areas of cortical GM which are known to cause the cognitive deficits typically linked to the clinical phenotype. Red arrows have been added to highlight focal areas of hypometabolism in each variant

lacunes, and perivascular spaces [76]. Whether these are separate processes or linked to the underlying disease cascade is actively being researched [77, 78]. There is evidence that they contribute equally and additively in individuals where no obvious impairment is present. In addition, individuals with heavy white matter burden tend to have more aggressive forms of the disease than those with limited or no signal of change in the white matter.

Advanced forms of MRI acquisitions are leading to better understanding of the diseases at different scales, from inferences made of the underlying tissue microstructure to how these forms of dementia disrupt the natural networks of the brain. Diffusion-

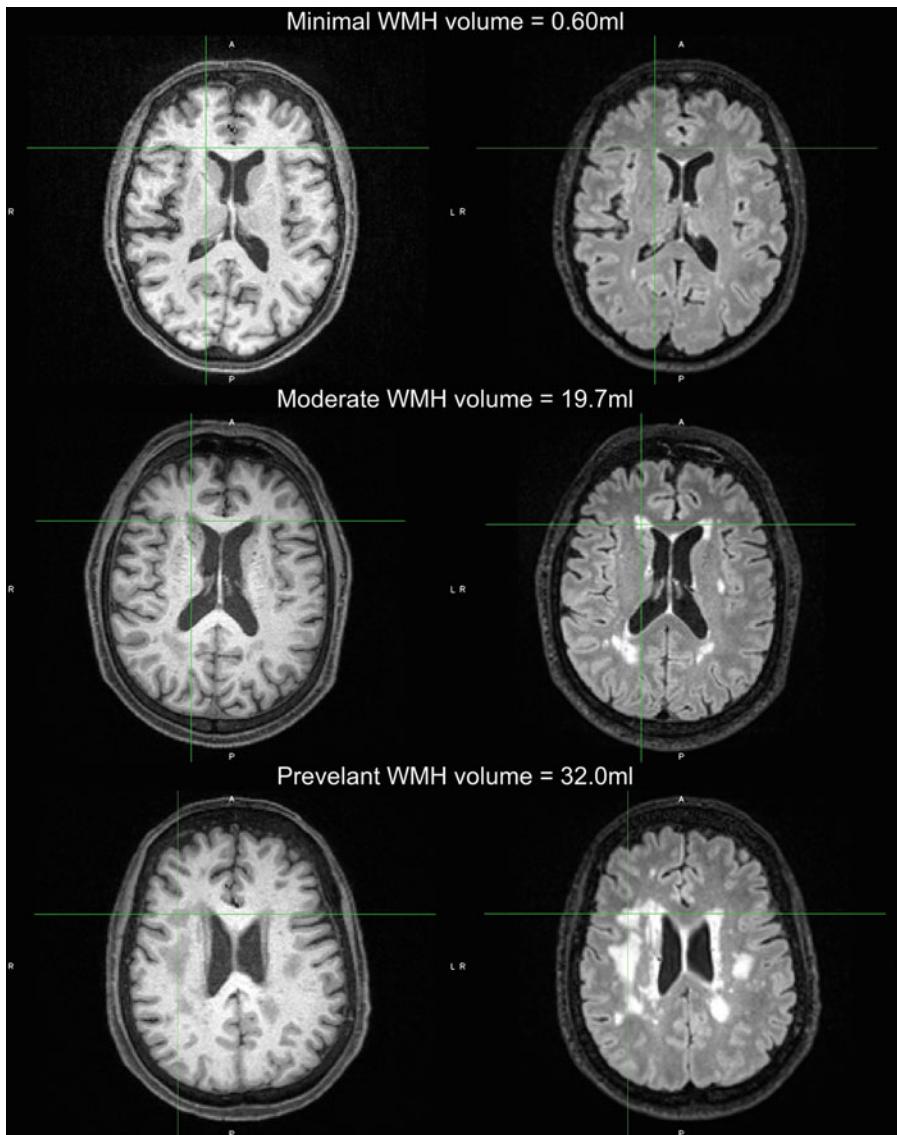


Fig. 4 Example T1-weighted (left column) and FLAIR (right column) images of minimal, moderate, and prevalent lesions in the white matter, often referred to as white matter hyperintensities (WMH), as they appear bright on FLAIR acquisitions. These lesions also often show up hypointense on T1-weighted scans, but FLAIR tends to be more sensitive and provide more contrast, particularly around deep gray matter areas

weighted imaging (DWI) provides measurements of both the magnitude and direction of the movement of water within a voxel. In white matter, the tissue consists of long fiber bundles that restrict the motion primarily along the direction of the fibers. In the cases of dementia, the integrity of these white matter bundles, whether through demyelination or some other form of neuronal dysfunction, tends to be less restrictive of water crossing boundaries, suggesting loss of microstructural integrity [79–82].

On a larger scale, connected brain networks can be identified with these techniques, either by tracing the diffusion profiles from one gray matter region to another using diffusion weighted imaging or by observing correlating patterns of deoxygenation of hemoglobin in the brain regions using functional MRI (fMRI) as a proxy for brain activity. These networks can become disrupted, primarily in regard to within network communication [83]. The direction of this disruption may depend on the stage of the disease. There is more evidence that the later stages of disease cause reduced connectivity and a disconnection from key seed regions to other areas. However, there may be an earlier stage where subjects compensate for increased pathology burden with hyperconnectivity [84, 85].

2.5 Advances in Novel Biomarkers

Novel assessments and biomarkers for all forms of dementia are highly active areas of research, from new fluid-based biomarkers to better computerized psychometric batteries to new imaging tracers and MR sequences to track additional aspects of the disease. While big data for machine learning in dementia has often meant assessing a large number of individuals, each with a small handful of measures, there are new forms of data collection that provide a rich set of data on single individuals. This could include not only the new epigenetic markers like single-cell RNA sequencing [86, 87] but also wearable devices that produce lots of data about individuals' daily activities and spatial navigation.

3 Challenges for Machine Learning

Researchers are nowadays focusing on two main aspects when it comes to AD and other forms of dementia. First, they aim to gain a better understanding of the disease process, including why individuals with similar underlying primary pathology result in different areas of the brain being affected, and thus have different clinical presentations. This is currently being investigated using many approaches ranging from molecular biology studies in wet labs to large epidemiological studies involving several thousands of participants. Second, they are developing tools to better assist clinicians with treatment management at the level of an individual. This includes, for example, the design of effective computational pipelines dedicated to patient diagnosis and prognosis.

Any research relying on machine learning methodologies must address the specific challenges presented by the disease. The first challenge comes from the large variability of diseases that makes it difficult to differentiate them, especially in the early stages. Additionally, mixed dementia, where patients have several diseases, is quite common. Indeed, the AD phenotype often coexists with vascular dementia or DLB. To partially address this issue, data from individuals with autosomal dominant forms of these diseases

are collected by international multicenter studies, as they often have earlier onset and usually “purer” forms of the disease. The Dominantly Inherited Alzheimer Network (DIAN)⁴ and the Genetic Frontotemporal Dementia Initiative (GENFI)⁵ are two studies collecting data from patients and relatives with familial AD and genetic FTD, respectively. While these studies have many benefits (*see Subheading 2.1*), there can be substantial differences between the genetic and the more widespread sporadic forms of these diseases, the most notable being younger disease onset and fewer comorbidities. Thus, there is a crucial need for ML methods that can disentangle the full complexity of sporadic forms of dementia. Finally, the variability comes not only from the presentation of the disease and comorbidities but also from the age at which the disease starts and the pace at which it progresses.

The second challenge is related to the duration of the disease, which often spans two decades and includes a yearslong prodromal phase. This makes it difficult to acquire data from individual patients that cover the full disease duration, especially as it is extremely challenging to identify with certainty who will develop the disease in the general population. While the previously mentioned studies of autosomal dominant forms of dementia can address this issue, it is not yet clear how much their findings can be translated to the far more common sporadic forms of these diseases. The Alzheimer’s Disease Neuroimaging Initiative (ADNI)⁶ is a large multicenter study that focuses on the acquisition of data from elderly individuals, consisting of those that are cognitively normal, those labeled as having mild cognitive impairment (MCI), and individuals diagnosed with probable AD [88]. These individuals are followed over several years, providing extremely valuable information for researchers. UK Biobank⁷ is another relevant initiative, as it aims to acquire in-depth phenotyping of half a million UK participants. Due to the large prevalence of AD and other related diseases in the population, it is anticipated that many individuals are in the pre-symptomatic phase of the diseases.

As aforementioned in the previous section, many markers of dementia are used to track the diseases, some being more relevant than others at specific times in the illness progression. For example, while amyloid PET-derived imaging biomarkers are valuable in the early stages of the disease, they are unable to quantify the progression toward the final stages. On the opposite end, clinical assessments, while being ineffective prior to symptomatic onset, enable monitoring of symptomatic evolution over time. This is a challenge

⁴ <https://dian.wustl.edu/>.

⁵ <https://www.genfi.org/>.

⁶ <https://adni.loni.usc.edu>.

⁷ <https://www.ukbiobank.ac.uk>.

as it requires to use the most relevant marker for the correct stage. In practice, this often leads to the use of complex models to handle large amount of multimodal data (imaging, clinical, genetic, demographic, ...). Additionally, each marker often suffers from its own variability, which can be intra-patient, inter-patient, and inter-center. For example, clinical assessments potentially differ based on the rater. MRI acquisitions will differ due to pulse sequence properties or scanner characteristics, as well as normal physiological variance such as hydration and caffeine intake, among others.

4 Machine Learning Developments

Machine learning has been used in multiple applications related to Alzheimer's disease and related dementia. As mentioned in the above section, dementia research has leveraged worldwide, multi-center studies in order to obtain enough data to characterize early changes and heterogeneity within the disease process. This has, in turn, propelled the development of dementia-focused ML applications. In this section, we review four main tasks in which extensive ML research has been performed.

Biomarker extraction from imaging data was originally done with manual assessments, which were time-consuming and subject to high inter-rater variability. Machine learning approaches that recreate these measurements with reduced time and variability have been a large effort that has served not only ADRD but many neurological disorders and neurodegenerative diseases. Given the numerous measurements that are now available on the datasets and the different aspects of the phenotype that they reflect, *disease classification and prediction* techniques have been used to identify consistent multivariate signatures between both healthy and disease groups, for differential diagnosis and for predicting the future state of patients. *Disease progression models* have been developed to determine the ordering of how markers go from normal to abnormal and to reconstruct the trajectories followed by these biomarkers, leading to advances in disease understanding and prognostication. *Data harmonization* to characterize variation caused by changes in scanning equipment and software across sites must be accounted for in order to obtain more accurate estimates of the biological changes.

4.1 Machine Learning-Derived Biomarkers

The largest area of machine learning research in relation to Alzheimer's disease and related disorders is to extract measurements from the different datasets. These biomarkers tend to reflect an aspect of function or integrity of the individual that will gain a better understanding of a disease. Changes in these biomarkers from normal values to abnormal provide a proxy for disease progression. Note that most of this research can usually be useful for other brain disorders.

Medical imaging provides valuable insights into an individual’s brain and is key to noninvasively assessing phenotypes due to the neurodegeneration process. Structural imaging, especially T1-weighted MRI, is commonly acquired in neurodegenerative studies as it enables quantifying key information such as atrophy in particular brain regions and thinning of the cortex or localized brain lesions. These features are often used as imaging biomarkers of disease progression, and many approaches relying on machine learning have been developed to extract them in the last three decades.

Brain segmentation and parcellation relate respectively to the classification of voxel into tissue types (e.g., gray matter, white matter, CSF) and the delineation of identified brain regions (e.g., whole brain, hippocampus, . . .).

The most popular open-source implementation (FSL⁸ [89], SPM⁹ [90]) for tissue segmentation relied on Gaussian mixture model optimized using expectation maximization [91, 92]. They enable to classify voxels based on their intensity but as well to accommodate with intensity inhomogeneity as well as noise via explicit modeling of the intensity bias field [91] and the use of Markov random field regularization, respectively [92]. With the advance of deep learning in the last decade, many techniques using convolutional neural networks have been proposed. Kumar et al. presented a U-Net based approach achieving close to 90% average Dice score coefficient on the segmentation for gray matter, white matter, and CSF on their dataset [93].

Classical approaches for brain parcellation rely on the concept of segmentation propagation and label fusion. In short, a set of template images, consisting of original images and associated labels, are aligned through medical image registration to a new image. The template labels are then warped into the shape of the new image’s brain and fused into a consensus segmentation. Popular approaches are HAMMER¹⁰ [94], FreeSurfer¹¹ [95], or Geodesic Information Flows (GIF) [96], among others. Using neural networks, de Brébisson et al. proposed a dedicated architecture concurrently using 2D and 3D patches and used iteratively to refine their results [97]. More recently, FastSurfer [98] was introduced, which is a deep learning-based method that aims to reproduce FreeSurfer’s results while considerably reducing processing time. While it is rapidly attracting users, further validation is needed to ensure that

⁸ <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki>.

⁹ <https://www.fil.ion.ucl.ac.uk/spm/>.

¹⁰ <https://www.nitrc.org/projects/hammer/>.

¹¹ <https://surfer.nmr.mgh.harvard.edu/>.

it is as robust as FreeSurfer across all disease types and severities. Finally, a current trend is to train from vast amounts of synthetic data with the hope of easing generalization to other sequences and/or resolutions. This is, for example, the approach taken in SynthSeg [99].

Other approaches have been proposed to segment individual regions of the brain that are particularly relevant to the study of dementia. For example, hippocampal segmentation has been an active area of research, with many studies including extensive validation in AD [100–104]. In the past years, the focus has turned to the segmentation of hippocampal subfields rather than the whole hippocampus [105, 106]. In particular, Manjon et al. used a U-Net approach combined with a deep supervision approach for training where their loss function optimizes segmentation accuracy at different image scales [107].

The identification of abnormalities (in particular those of vascular origin) is also a key step in the study of dementia and particularly for differential diagnosis. These abnormalities include hyper- or hypo-intensity lesions, micro-bleeds, perivascular spaces, or lacunes. Various approaches to segment T2/FLAIR white matter hyperintensities have been proposed (see [108] for a comparison of seven of them), while there have been fewer works on micro-bleeds or lacunes. Sudre et al. proposed a Gaussian mixture model approach with automated detection of classes number to accurately segment brain tissue classes, as well as abnormalities [109]. More recently, deep learning approaches have also been developed. For example, Boutinaud et al. used a U-Net, which parameters were pre-trained using an autoencoder, to automatically segment perivascular spaces from T1-weighted MRI scans [110]. Wu et al. also used a U-Net architecture for segmentation hyperintensities from T1-weighted and FLAIR MR images [111].

4.2 Disease Classification and Prediction

Machine learning is a powerful tool when it comes to disease diagnosis and prognosis. As a result, many approaches have been proposed for disease classifications, to identify the current stage of a disease within an individual or to predict their future state (e.g., transition to dementia in patients with MCI).

For over a decade, dozens (if not hundreds) of papers have proposed classification techniques to distinguish patients diagnosed with AD versus age-matched controls (e.g., [112–119]) or patients suffering from mild cognitive impairment who are staying stable in time versus those who will progress to a diagnosis of AD (e.g., [115, 117–124]). The latter task can contribute to prognosis which, when it comes to dementia-related diseases, often consists of classifying patients who are likely going to convert from mild cognitive impairment to symptomatic AD within a given time

interval, typically 3 years,¹² from those who are going to stay stable. Several literature reviews on dementia classification and prediction have been published [125–130]. In particular, recent reviews by Jo et al. [126] and Ansart et al. [130] have covered the topic of prognosis.

The proposed methodologies have been extremely varied not only in terms of ML algorithms but also of input modalities and extracted features. Initial ML methods often used support vector machines (e.g., [112, 114]), while subsequent works used more recent techniques such as random forest [117] and Gaussian processes [121, 131]. Many recent works have used deep learning classification techniques [127, 129]. However, so far, deep learning has not outperformed classical ML for AD classification and prediction [129, 130, 132]. Furthermore, a review on convolutional neural networks for AD diagnosis from T1-weighted MRI [129] has identified that more than half of these deep learning studies may have been contaminated by data leakage, which is particularly worrisome. In terms of features, some studies use the whole brain as input, either using directly the raw image or computing voxel-wise (or vertex-wise when considering the cortical surface) measures [125]. Others parcellate the brain into regions of interest, within which features are computed. In particular, researchers have combined segmentation approaches with disease classification techniques. This has the advantage of limiting the search space of the machine learning approach via the use of prior knowledge. For example, Coupe et al. used a patch-based approach to classify voxels from the hippocampus as it is known to be a vulnerable structure in patients with dementia [116]. The input modalities have also been extremely varied. While earlier works often focused on T1-weighted MRI only [112–114], subsequent studies have included other imaging modalities, in particular FDG-PET [121, 123]. Other researchers have combined tailored features extracted from images and non-images features such as fluid biomarkers [121], cognitive tests [124, 133], APOE genotype [121], or genome-wide genotyping data [134, 135]. Through deep learning, researchers are avoiding the need to craft features and can use traditional deep learning approaches to directly infer disease status from raw data: imaging or non-imaging. However, to date, there has been less interest in this area than in biomarker extraction, and thus fewer innovative solutions have been proposed. Popular architectures include conventional neural network, autoencoder, and recurrent neural networks, among others [127]. Training strategies mostly relied on supervised approaches, where some groups have relied on pre-trained networks to compensate for relatively

¹² While 3 years is certainly a relevant time frame to provide useful information to patients and relatives, it is likely that the focus of research on such time frame was largely driven by the typical follow-up which is available for most patients in large publicly available databases such as ADNI.

small training databases. However, as mentioned above, it has not been demonstrated so far that deep learning outperforms conventional ML for dementia classification.

A large portion of the literature has relied on the ADNI database. While having such a rich and large database has propelled the development of algorithms, one can wonder if the proposed method will generalize well to other datasets, a problem which has less often been addressed. Another worrisome aspect is that many of the papers based on ADNI are difficult to reproduce because they lack a description of the subjects used and because the code is not often available [136]. They are also difficult to compare. In particular, since different preprocessing tools are used, it is often difficult to know whether improvements in performance come from the innovation in ML or from the preprocessing. Standardized datasets have been created in ADNI [137] to address the former issue. Whenever possible, one of these datasets should be used, or authors should provide a list of subjects/scans included in the study for the purpose of reproducibility.

Researchers have organized challenges that can provide an objective comparison of algorithms. One can cite in particular the CADDementia challenge for classification [138] and the TADPOLE challenge for prognosis [132]. Such challenges provide very important and useful information on the respective merits of different approaches. However, more challenges, in particular using more diverse data, would be needed.

To a lesser extent, differential diagnosis has also been addressed. Earlier works focused on classifying patients diagnosed with AD versus patients diagnosed with frontotemporal dementia [112, 139]. More recent studies have considered classifying between various types of dementia [140, 141].

Overall, there has been considerable research in AD classification and prediction. The easiest task, which is the classification of patients diagnosed with AD versus normal controls, can be considered as solved with accuracy typically above 90%, at least when using data which quality is comparable to that of research studies. However, this task has little clinical utility. For the more interesting task of prediction progression to AD in MCI patients, the performance has increased over the years, with AUC now above 80% [130]. Interestingly, it has been shown that studies which include cognitive tests and FDG-PET tend to have better results than those using T1-weighted MRI only [130]. It is particularly noteworthy that cognitive tests tend to be overlooked, given that they are relatively cheap to perform and widely available. This probably reflects the fact that much of these works have arisen in the medical image computing community. Finally, other tasks are still short of becoming clinically useful, such as the development of a multi-pathology differential diagnostic classifier. This is possibly due to the lack of tailored methodologies leveraging all available information. This will be further discussed in the conclusions of this chapter.

4.3 Disease Progression Modeling

Following the release of a hypothetical model of disease progression by Jack et al. [142], researchers have tried to create data-driven models that would accurately describe the various components of the underlying disease process. These so-called *disease progression models* have been developed to understand the ordering of phenotypic events (such as a marker becoming abnormal), to model the variability of ordering or trajectories within a population and to be able to distinguish between different disease subtypes. The reader may refer to Chap. 17 for a detailed description of the methodology underlying data-driven disease progression models. Event-based models (EBM) [143] have been created to learn from a curated population and across different modalities the different hidden states of a disease. EBMs are able to order all input features from the one that will most likely become abnormal first to the one that becomes abnormal last [143]. The application of EBM in dementia has been very successful, including studies in familial AD [143, 144], sporadic AD [145, 146], posterior cortical atrophy [147], and genetic FTD [148]. An extended approach called Subtype and Stage Inference (SuStaIn) was developed by Young et al., which incorporates clustering to characterize disease subtypes. In particular, it allowed uncovering the symptomatic profiles of different variants of genetic FTD as well as AD subtypes [29]. EBMs have the advantage of being applicable to cross-sectional data but only provide an ordering of events with no temporal scale as to when they become abnormal. In most EBMs, there is also an assumption of a monotonic biomarker trajectory, an assumption which has been questioned in the early stages of AD [149]. Other works have leveraged longitudinal data to build continuous trajectories. Jedynak et al. used a disease progression model to derive a progression score on a linear scale for every individual in AD [150]. Schiratti et al. proposed a general nonlinear mixed effects model that can handle not only scalar biomarker data but also images or shapes [151]. Applied to AD, the approach uncovered trajectories of progression for different variables, including cognitive tests, PET-derived hypometabolism, and local hippocampal atrophy [152]. In 2021, Wijeratne and Alexander proposed an approach that can, using longitudinal data, infer both discrete event ordering (as in EBMs) and continuous trajectories [153]. Lastly, Abi Nader et al. proposed SimulAD, which enables the setup of *in silico* interventional trial, where patient prognosis can be assessed against several possible therapies (drug type or timing of intervention) [154].

4.4 Data Harmonization

Data heterogeneity inducing bias arises from many sources, including differences in acquisition protocols, acquisition devices, or populations under study. This is true in large multicenter studies such as ADNI, DIAN, and GENFI. However, such research studies use harmonized protocols for data acquisition and perform strict

quality control. It becomes even worse when one is using clinical routine data which is not acquired with harmonized protocols and can be of extremely varied quality.

Data heterogeneity may have various impacts on ML algorithms. For example, for any classification task, one wants to ensure the model focuses on the diseases' features rather than on any differences caused by the acquisition sites. The ability to harmonize data from any origin is also critical to translate any analytical tool in clinical practice where acquisition protocols are rarely standardized.

This is especially true for imaging biomarkers where scans often contain so-called scanner signatures. As a result, images are often preprocessed prior to being used with machine learning algorithms. The preprocessing steps involve intensity normalization in the shape of image filtering, for denoising, or intensity histogram normalization. For example, Erus et al. [155] proposed a framework to achieve consistent segmentation of brain structures across multiple sites. Their approach relies on the creation of site-specific atlases while ensuring consistency between all available atlases. Their evaluation shows that they reduce the variability associated with sites on volumetric measurements, key to track the process of brain atrophy, derived from structural images. Another example is the work of Jog et al. [156], who used image synthesis via contrast learning to harmonize images acquired with different pulse sequences. The Removal of Artificial Voxel Effect by Linear regression (RAVEL) approach by Fortin et al. [157] is another exemplar application of data harmonization. It consists of a voxel-wise intensity normalization technique, where they apply singular value decomposition (SVD) of the control voxels to estimate factors of unwanted variation. The control voxels are those unaffected by the pathology, such as those in the cerebrospinal fluid (CSF). The unwanted factors are then estimated using linear regression for every voxel of the brain, and the residuals are taken as the RAVEL-corrected intensities. This model has then been further extended [158] to include the modeling of site-specific scaling factors on summary measures derived from the images. Using empirical Bayes to improve the estimation of the site, this model can be used to correct several imaging modalities while associating relevant clinical and demographic information. It was originally developed to correct gene expression microarray data [159], being later extended to correct DTI maps [158], cortical thickness measurements [160], or structural MRI [161]. Additional extensions include longitudinal data [162], site effects due to covariance [163], and a generalized additive model in order to handle nonlinear trajectories over the life span [164]. Prado et al. [165] proposed Dementia ConnEEGtom to harmonize neurophysiological data. They propose a whole analytical pipeline involving many steps, including denoising, artifact removal, and spatial normalization, to promote a standardized

processing of EEG data, thereby enabling their use in machine learning while minimizing the bias that could be induced by variability in data handling across sites.

Acquired clinical scales also need to be standardized and harmonized between centers, especially when used jointly in a single machine learning approach. Costa et al. [166] provide recommendations on how this should be achieved and which scales should be acquired for the neuropsychological assessment in neurodegenerative diseases. Even when using the same scales across different multicenter studies, it is important to understand that the provenance and contextual information of each study must be considered, since that might introduce bias in the training of the models [167].

5 What Is Next?

This chapter has illustrated that dementia is a complex, multifactorial, heterogeneous set of pathologies and syndromes, sometimes occurring in parallel. As a result, a wide variety of clinical, genetic, cognitive, imaging, and biofluid data have been collected to characterize these disease processes over a large number of different cohorts, from both those individuals suffering from various forms of dementia, as well as those at risk of developing a form of dementia due to their genetic/environmental/pathophysiological risk profile. Despite the wealth of data on dementia that is available to machine learning researchers, there are still limitations, both in terms of the data available and in the current thinking about how to apply machine learning, that must be addressed in order for machine learning to reach its true potential in terms of making an impact on these conditions. Key to this validation will be open science initiatives that allow for reproducibility and replication of results, so that the value can be demonstrated in independent cohorts.

Careful thought must be given to how to incorporate the myriad different data types that are available to researchers. Despite the wealth of multimodality information and big data available, much of the machine learning-based research in dementia has only considered a subset of the available information. This is the case even within a single community such as the imaging one where each modality or pulse sequence is most often analyzed individually. Machine learning approaches are however able to extract highly nonlinear information, which should enable the development of truly multi-data frameworks able to capture the complexity of the diseases. At the same time, when multiple features across different data acquisition domains are combined into a single analysis, particularly in those individuals who already show evidence of impairment, there is a higher likelihood of missing or corrupted

information. The simplest solution to the problem of missing data, and often the most implemented, is to perform complete case analyses, discarding any observations where any of the variables are missing. This practice results in the time donated by patients, carers, and volunteers willing to support research efforts through extensive and often onerous data collection being squandered and the full potential of the resulting data not being realized. While it is common to have 30% of data being discarded in some large studies, we could benefit hugely from further research in this direction. Multiple imputation techniques could be used to address this issue and ensure all available data can be used.

A related question is how to address the limitations of cross-sectional snapshots of data in a decades-long disease process and how best to use relatively short-term follow-up data. While inference on cross-sectional measures would be ideal in terms of providing information expediently to a patient, there are many confounds and covariates that can contribute to added variability, making classification tasks, particularly in the early stages of the disease, less accurate. Longitudinal data, particularly in imaging modalities, tends to reduce the influence of these confounds, and the within-subject change can be more sensitive to identifying the prognosis of individual trajectories. However, requiring longitudinal data for a classification task is undesirable for patients, who would be provided with no information until they come back for additional testing in a year or two. Thus, machine learning approaches could investigate whether a hybrid approach might be more powerful: triaging first with the cross-sectional data and only requiring longitudinal data in cases where inference cannot be made at the baseline assessment with a high degree of confidence.

A second challenge is to extend machine learning approaches to datasets that are more reflective of standard clinical settings. This refers not only to the type of data that is collected but also to the conditions under which the data is collected and to the populations within which data is acquired. Clinical research studies conducted at research institutions often include advanced data acquisitions that are costly and time-consuming, making them intractable for translation into wider communities and developing countries. There is thus a mismatch between the quality of the data acquired in research settings compared to the data acquired in day-to-day clinical environments. For example, only a subset of the patients suspected of having dementia undergo medical imaging and from this subset only a fraction of these individuals are offered MRI scans. In the majority of cases, CT are acquired, and they are mostly used to rule out other causes for impairment, such as space-occupying lesions. As a result, a large amount of CT scans are collected, and they could potentially be an important resource to develop computer-assisted tools able to reach a larger population [168]. Even when the same data type is acquired in clinical and

research settings, there can be a considerable mismatch in terms of data quality and homogeneity. For instance, clinical routine MRI is of extremely variable quality and usually acquired using non-harmonized protocols. Similarly, with the current democratization of wearable and smartphone collection data which is prime to be processed with machine learning, there are opportunities to develop novel frameworks assisting patients, carers, and clinicians.^{13,14} Another aspect of this challenge to improve widespread translation is that clinical research studies typically involve a disproportionate amount of affluent Caucasian individuals of European descent, meaning that we do not yet have enough data to fully quantify the heterogeneity that is observed in other ethnic groups. While large-scale cohort studies are looking to address this issue, focusing on assembling appropriate testing and training sets to represent the diversity of the population will be an important element for improving machine learning performance in the future.

Effective partnerships with the clinicians who are using the data are another key challenge in terms of incorporating machine learning in a wider clinical setting [169, 170]. Clinicians must believe in the added value that novel machine learning approaches can provide in order to incorporate them as part of their clinical workup and decision-making. One approach to achieving this buy-in from clinicians is a push toward “explainable AI,” such that the results from machine learning algorithms make intuitive sense and that the clinician can better understand how the algorithm came to that decision. While there are certainly concerns around the opaqueness of some algorithms that could lead to overfitting or spurious results, an insistence on explainable AI may also restrict the development of better algorithms that can provide more value, and in some cases, it may result in reducing a complex multivariate pattern down to a summary measure that can be understood, throwing away valuable information in the process. What is likely more important is that best practices are followed in terms of training, testing, model development, and validation of the algorithm such that clinicians may not necessarily understand how the algorithm achieved a specific result but that they are convinced by the evidence of the value it provides.

The final, and likely most significant, challenge is how best to characterize heterogeneity and mixed pathology. Classification of well-characterized cohorts of clear cases of AD and normal controls provides little benefit, in particular given the rise of accurate and increasingly cheaper and accessible plasma tests. Dementia covers a wide range of symptoms caused by a myriad of pathologies and etiologies occurring in a decades-long process, and correctly

¹³ <https://edon-initiative.org>.

¹⁴ <https://www.ftdtalk.org/research/our-projects/digital/>.

identifying the underlying disease will be critical for treatment plans as disease-modifying therapies become available. There is a natural inclination to thus subdivide and characterize a number of distinct and discrete disorders, which has led to a focus on machine learning algorithms to aid in tasks of differential diagnosis. However, there are often no clear boundaries between the phenotypic profiles of these disorders, and mixed pathologies are common. Therefore, there should be a shift in machine learning away from classifying between normal aging and a single disease or differential diagnosis task with the goal of dichotomizing between two or more disorders but rather a probabilistic framework that allows for multiple pathologies to coexist. With the advance of big data analysis, access to clinical care data, and innovative machine learning, all is in place for this shift to be achieved.

6 Conclusion

Dementia of all forms is going to be one of the biggest global health challenges around the globe over the coming decades. Improving the ability to characterize the disease at an early stage and providing an accurate prognosis that allows doctors to provide effective treatment plans and individuals to make informed decisions about how to manage their affairs are going to be critical in order to reduce the distress and burden experienced by people suffering from these diseases and their families. Machine learning will need to play a key role in achieving these targets.

Acknowledgements

MM, LC, and SO research is supported by the Wellcome EPSRC Centre for Medical Engineering at King's College London (WT 203148/Z/16/Z), EPSRC (EP/T022205/1), the Wellcome Trust (WT 215010/Z/18/Z), the UK Research and Innovation London Medical Imaging and Artificial Intelligence Centre for Value-Based Healthcare, Medical Research Council (MRC), NIHR, Alzheimer's Society, and the European Union.

DMC is supported by the UK Dementia Research Institute which receives its funding from DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK (ARUK-PG2017-1946), the UCL/UCLH NIHR Biomedical Research Centre, and the UKRI Innovation Scholars: Data Science Training in Health and Bioscience (MR/V03863X/1).

MB is supported by a Fellowship award from the Alzheimer's Society, UK (AS-JF-19a-004-517). MB's work was also supported by the UK Dementia Research Institute which receives its funding

from DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society, and Alzheimer's Research UK.

We would like to thank Dr. Jonathan Schott from the UCL Queen Square Institute of Neurology, Avid Radiopharmaceuticals, and the NIHR Biomedical Research Centre for the images used in the figures. We also would like to thank Prof. Nick Fox for kindly supporting all of our dementia-related research projects throughout the years.

References

- Gauthier S, Rosa-Neto P, Morais JA, Webster C (2021) World Alzheimer report 2021: journey through the diagnosis of dementia. *Alzheimer's Disease International*. <https://www.alzint.org/resource/world-alzheimer-report-2021/>
- McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984) Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group under the auspices of department of health and human services task force on Alzheimer's disease. *Neurology* 34:939–944. <https://pubmed.ncbi.nlm.nih.gov/6610841/>
- McKhann GM, Knopman DS, Chertkow H, Hyman BT, Clifford RJ Jr, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (2011) The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association work-groups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dementia* 7:263–269. <https://doi.org/10.1016/J.JALZ.2011.03.005>
- Jack CR, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, Thies B, Phelps CH (2011) Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association work-groups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dementia* 7:257–262. <https://doi.org/10.1016/J.JALZ.2011.03.004>
- Ryan NS, Nicholas JM, Weston PSJ, Liang Y, Lashley T, Guerreiro R, Adamson G, Kenny J, Beck J, Chavez-Gutierrez L, de Strooper B, Revesz T, Holton J, Mead S, Rossor MN, Fox NC (2016) Clinical phenotype and genetic associations in autosomal dominant familial Alzheimer's disease: a case series. *Lancet Neurol* 15:1326–1335. [https://doi.org/10.1016/S1474-4422\(16\)30193-4](https://doi.org/10.1016/S1474-4422(16)30193-4)
- Ryman DC, Acosta-Baena N, Aisen PS, Bird T, Danek A, Fox NC, Goate A, Frommelt P, Ghetti B, Langbaum JBS, Lopera F, Martins R, Masters CL, Mayeux RP, McDade E, Moreno S, Reiman EM, Ringman JM, Salloway S, Schofield PR, Sperling R, Tarot PN, Xiong C, Morris JC, Bateman RJ (2014) Symptom onset in autosomal dominant Alzheimer disease: a systematic review and meta-analysis. *Neurology* 83: 253–260. <https://doi.org/10.1212/WNL.0000000000000596>
- Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N, van Duijn CM (1997) Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis. *JAMA* 278(16):1349–1356. <https://doi.org/10.1001/jama.1997.03550160069041>
- Genin E, Hannequin D, Wallon D, Sleegers K, Hiltunen M, Combarros O, Bulldido MJ, Engelborghs S, Deyn PD, Berr C, Pasquier F, Dubois B, Tognoni G, Fiévet N, Brouwers N, Bettens K, Arosio B, Coto E, Zompo MD, Mateo I, Epelbaum J, Frank-Garcia A, Helisalmi S, Porcellini E, Pilotto A, Forti P, Ferri R, Scarpini E, Siciliano G, Solfrizzi V, Sorbi S, Spalletta G, Valdivieso F, Vepsäläinen S, Alvarez V, Bosco P, Mancuso M, Panza F, Nacmias B, Boss P, Hanon O, Piccardi P, Annoni G, Seripa D, Galimberti D, Licastro F, Soininen H, Dartigues JF, Kamboh MI, Broeckhoven CV, Lambert JC, Amouyel P, Campion D (2011) APOE and Alzheimer disease: a major gene with semi-dominant inheritance. *Mol Psychiatry* 16:903–907. <https://doi.org/10.1038/mp.2011.52>
- Crutch SJ, Schott JM, Rabinovici GD, Murray M, Snowden JS, Flier WM, Dickerson BC, Vandenberghe R, Ahmed S, Bak TH, Boeve BF, Butler C, Cappa SF, Ceccaldi M, Souza LC, Dubois B, Felician O, Galasko D, Graff-Radford J, Graff-Radford NR, Hof PR,

- Krolak-Salmon P, Lehmann M, Magnin E, Mendez MF, Nestor PJ, Onyike CU, Pelak VS, Pijnenburg Y, Pramstötz S, Rossor MN, Ryan NS, Scheltens P, Shakespeare TJ, González AS, Tang-Wai DF, Yong KX, Carrillo M, Fox NC (2017) Consensus classification of posterior cortical atrophy. *Alzheimer's Dementia* 13:870–884. <https://doi.org/10.1016/j.jalz.2017.01.014>
10. Gorno-Tempini ML, Hillis AE, Weintraub S, Kertesz A, Mendez M, Cappa SF, Ogar JM, Rohrer JD, Black S, Boeve BF, Manes F, Drionkers NF, Vandenberghe R, Rascovsky K, Patterson K, Miller BL, Knopman DS, Hodges JR, Mesulam MM, Grossman M (2011) Classification of primary progressive aphasia and its variants. *Neurology* 76:1006–1014. <https://doi.org/10.1212/WNL.0B013E31821103E6>
11. Nelson PT, Dickson DW, Trojanowski JQ, Jack CR, Boyle PA, Arfanakis K, Rademakers R, Alfuzoor I, Attems J, Brayne C, Coyle-Gilchrist ITS, Chui HC, Fardo DW, Flanagan ME, Halliday G, Hukkanen SRK, Hunter S, Jicha GA, Katsumata Y, Kawas CH, Keene CD, Kovacs GG, Kukull WA, Levey AI, Makkinejad N, Montine TJ, Murayama S, Murray ME, Nag S, Rissman RA, Seeley WW, Sperling RA, III CLW, Yu L, Schneider JA (2019) Limbic-predominant age-related TDP-43 encephalopathy (LATE): consensus working group report. *Brain* 142:1503–1527. <https://doi.org/10.1093/brain/awz099>
12. Román GC, Tatemichi TK, Erkinjuntti T, Cummings JL, Masdeu JC, Garcia JH, Amaducci L, Orgogozo JM, Brun A, Hofman A, Moody DM, O'Brien MD, Yamaguchi T, Grafman J, Drayer BP, Bennett DA, Fisher M, Ogata J, Kokmen E, Bermejo F, Wolf PA, Gorelick PB, Bick KL, Pajewski AK, Bell MA, Decarli C, Culebras A, Korczyn AD, Bogousslavsky J, Hartmann A, Scheinberg P (1993) Vascular dementia: diagnostic criteria for research studies. report of the NINDS-AIREN international workshop. *Neurology* 43:250–260. <https://doi.org/10.1212/WNL.43.2.250>
13. Joutel A, Vahedi K, Corpechot C, Troesch A, Chabriat H, Vayssiére C, Cruaud C, Maciazeck J, Weissenbach J, Bousser MG, Bach JF, Tournier-Lasserre E (1997) Strong clustering and stereotyped nature of Notch3 mutations in CADASIL patients. *Lancet* 350: 1511–1515. [https://doi.org/10.1016/S0140-6736\(97\)08083-5](https://doi.org/10.1016/S0140-6736(97)08083-5)
14. Warren JD, Rohrer JD, Rossor MN (2013) Frontotemporal dementia. *BMJ* 347. <https://doi.org/10.1136/BMJ.F4827>
15. Moore KM, Nicholas J, Grossman M, McMillan CT, Irwin DJ, Massimo L, Deerlin VMV, Warren JD, Fox NC et al (2020) Age at symptom onset and death and disease duration in genetic frontotemporal dementia: an international retrospective cohort study. *Lancet Neurol* 19:145–156. [https://doi.org/10.1016/S1474-4422\(19\)30394-1](https://doi.org/10.1016/S1474-4422(19)30394-1)
16. Rascovsky K, Hodges JR, Knopman D, Mendez MF, Kramer JH, Neuhaus J, Swieten JCV, Seelaar H, Doppler EG, Onyike CU, Hillis AE, Josephs KA et al (2011) Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain* 134:2456–2477. <https://doi.org/10.1093/BRAIN/AWR179>
17. Mackenzie IR, Neumann M (2016) Molecular neuropathology of frontotemporal dementia: insights into disease mechanisms from postmortem studies. *J Neurochem* 138:54–70. <https://doi.org/10.1111/JNC.13588>
18. Lashley T, Rohrer JD, Mead S, Revesz T (2015) Review: an update on clinical, genetic and pathological aspects of frontotemporal lobar degenerations. *Neuropathol Appl Neuropathol* 41:858–881. <https://doi.org/10.1111/NAN.12250>
19. Whitwell JL, Przybelski SA, Weigand SD, Ivnik RJ, Vemuri P, Gunter JL, Senjem ML, Shiung MM, Boeve BF, Knopman DS, Parisi JE, Dickson DW, Petersen RC, Jack CR, Josephs KA (2009) Distinct anatomical subtypes of the behavioural variant of frontotemporal dementia: a cluster analysis study. *Brain* 132:2932–2946. <https://doi.org/10.1093/BRAIN/AWP232>
20. Gordon E, Rohrer JD, Fox NC (2016) Advances in neuroimaging in frontotemporal dementia. *J Neurochem* 138:193–210. <https://doi.org/10.1111/JNC.13656>
21. Bocchetta M, Malpetti M, Todd EG, Rowe JB, Rohrer JD (2021) Looking beneath the surface: the importance of subcortical structures in frontotemporal dementia. *Brain Commun* 3. <https://doi.org/10.1093/BRAINCOMMS/FCAB158>
22. Rohrer JD, Warren JD, Modat M, Ridgway GR, Douiri A, Rossor MN, Ourselin S, Fox NC (2009) Patterns of cortical thinning in the language variants of frontotemporal lobar degeneration. *Neurology* 72:1562–1569. <https://doi.org/10.1212/WNL.0B013E3181A4124E>

23. Rohrer JD, Lashley T, Schott JM, Warren JE, Mead S, Isaacs AM, Beck J, Hardy J, Silva RD, Warrington E, Troakes C, Al-Sarraj S, King A, Borroni B, Clarkson MJ, Ourselin S, Holton JL, Fox NC, Revesz T, Rossor MN, Warren JD (2011) Clinical and neuroanatomical signatures of tissue pathology in frontotemporal lobar degeneration. *Brain* 134:2565–2581. <https://doi.org/10.1093/BRAIN/AWR198>
24. Woollacott IO, Rohrer JD (2016) The clinical spectrum of sporadic and familial forms of frontotemporal dementia. *J Neurochem* 138: 6–31. <https://doi.org/10.1111/JNC.13654>
25. Rohrer JD, Nicholas JM, Cash DM, van Swieten J, Doppler E, Jiskoot L, van Minkelen R, a Rombouts S, Cardoso MJ, Clegg S, Espak M, Mead S, Thomas DL, Vita ED, Masellis M, Black SE, Freedman M, Keren R, MacIntosh BJ, Rogaeva E, Tang-Wai D, Tartaglia MC, Laforce R, Tagliavini F, Tiraboschi P, Redaelli V, Prioni S, Grisoli M, Borroni B, Padovani A, Galimberti D, Scarpini E, Arighi A, Fumagalli G, Rowe JB, Coyle-Gilchrist I, Graff C, Fallström M, Jelic V, Ståhlbom AK, Andersson C, Thonberg H, Lilius L, Frisoni GB, Binetti G, Pievani M, Bocchetta M, Benussi L, Ghidoni R, Finger E, Sorbi S, Nacmias B, Lombardi G, Polito C, Warren JD, Ourselin S, Fox NC, Rossor MN (2015) Presymptomatic cognitive and neuroanatomical changes in genetic frontotemporal dementia in the Genetic Frontotemporal dementia Initiative (GENFI) study: a cross-sectional analysis. *Lancet Neurol* 14: 253–262. [https://doi.org/10.1016/S1474-4422\(14\)70324-2](https://doi.org/10.1016/S1474-4422(14)70324-2)
26. Cash DM, Bocchetta M, Thomas DL, Dick KM, van Swieten JC, Borroni B, Galimberti D, Masellis M, Tartaglia MC, Rowe JB, Graff C, Tagliavini F, Frisoni GB, Laforce RJ, Finger E, de Mendonca A, Sorbi S, Rossor MN, Ourselin S, Rohrer JD (2018) Patterns of grey matter atrophy in genetic frontotemporal dementia: results from the genfi study. *Neurobiol Aging* 62: 191–196. <https://doi.org/10.1016/j.neurobiolaging.2017.10.008>
27. Whitwell JL, Weigand SD, Boeve BF, Senjem ML, Gunter JL, DeJesus-Hernandez M, Rutherford NJ, Baker M, Knopman DS, Wszolek ZK, Parisi JE, Dickson DW, Petersen RC, Rademakers R, Jack CR, Josephs KA (2012) Neuroimaging signatures of frontotemporal dementia genetics: C9orf72, tau, progranulin and sporadics. *Brain* 135:794. <https://doi.org/10.1093/BRAIN/AWS001>
28. Sha SJ, Takada LT, Rankin KP, Yokoyama JS, Rutherford NJ, Fong JC, Khan B, Karydas A, Baker MC, DeJesus-Hernandez M, Sha SJ, Takada LT, Rankin KP, Yokoyama JS, Rutherford NJ, Fong JC, Khan B, Karydas A, Baker MC, DeJesus-Hernandez M, Pribadi M, Coppola G, Geschwind DH, Rademakers R, Lee SE, Seeley W, Miller BL, Boxer AL (2012) Frontotemporal dementia due to c9orf72 mutations: clinical and imaging features. *Neurology* 79:1002–1011. <https://doi.org/10.1212/WNL.0b013e318268452e>
29. Young AL, Marinescu RV, Oxtoby NP, Bocchetta M, Yong K, Firth NC, Cash DM, Thomas DL, Dick KM, Cardoso J (2018) Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nat Commun* 9:4273
30. Young AL, Bocchetta M, Russell LL, Convery RS, Peakman G, Todd E, Cash DM, Greaves CV, van Swieten J, Jiskoot L, Seelaar H, Moreno F, Sanchez-Valle R, Borroni B, Laforce R, Masellis M, Tartaglia MC, Graff C, Galimberti D, Rowe JB, Finger E, Synofzik M, Vandenberghe R, de Mendonça A, Tagliavini F, Santana I, Ducharme S, Butler C, Gerhard A, Levin J, Danek A, Otto M, Sorbi S, Williams SC, Alexander DC, Rohrer JD (2021) Characterizing the clinical features and atrophy patterns of mapt-related frontotemporal dementia with disease progression modeling. *Neurology* 97: e941–e952. <https://doi.org/10.1212/WNL.0000000000012410>
31. McKeith IG, Boeve BF, DIckson DW, Halliday G, Taylor JP, Weintraub D, Aarsland D, Galvin J, Attems J, Ballard CG, Bayston A, Beach TG, Blanc F, Bohnen N, Bonanni L, Bras J, Brundin P, Burn D, Chen-Plotkin A, Duda JE, El-Agnaf O, Feldman H, Ferman TJ, Ffytche D, Fujishiro H, Galasko D, Goldman JG, Gomperts SN, Graff-Radford NR, Honig LS, Iranzo A, Kantarci K, Kaufer D, Kukull W, Lee VM, Leverenz JB, Lewis S, Lippa C, Lunde A, Masellis M, Masliah E, McLean P, Mollenhauer B, Montine TJ, Moreno E, Mori E, Murray M, O'Brien JT, Orimo S, Postuma RB, Ramaswamy S, Ross OA, Salmon DP, Singleton A, Taylor A, Thomas A, Tiraboschi P, Toledo JB, Trojanowski JQ, Tsuang D, Walker Z, Yamada M, Kosaka K (2017) Diagnosis and management of dementia with Lewy bodies. *Neurology* 89: 88–100. <https://doi.org/10.1212/WNL.000000000004058>

32. Rongve A, Aarsland D (2020) The Lewy body dementias: dementia with Lewy bodies and Parkinson's disease dementia. In: Oxford textbook of old age psychiatry, pp 495–512. <https://doi.org/10.1093/MED/9780198807292.003.0032>
33. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A, Pedersen NL (2006) Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry* 63:168–174. <https://doi.org/10.1001/ARCHPSYC.63.2.168>
34. Selkoe DJ, Hardy J (2016) The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol Med* 8:595–608. <https://doi.org/10.15252/emmm.201606210>
35. Borroni B, Ferrari F, Galimberti D, Nacmias B, Barone C, Bagnoli S, Fenoglio C, Piaceri I, Archetti S, Bonvicini C, Gennarelli M, Turla M, Scarpini E, Sorbi S, Padovani A (2014) Heterozygous TREM2 mutations in frontotemporal dementia. *Neurobiol Aging* 35:934.e7–934.e10. <https://doi.org/10.1016/J.NEUROBIOLAGING.2013.09.017>
36. Chasioti D, Yan J, Nho K, Saykin AJ (2019) Progress in polygenic composite scores in Alzheimer's and other complex diseases. *Trends Genet* 35:371. <https://doi.org/10.1016/J.TIG.2019.02.005>
37. Altmann A, Scelsi MA, Shoaib M, Silva ED, Aksman LM, Cash DM, Hardy J, Schott JM (2020) A comprehensive analysis of methods for assessing polygenic burden on Alzheimer's disease pathology and risk beyond APOE. *Brain Commun* 2. <https://doi.org/10.1093/BRAINCOMMS/FCZ047>
38. Stocker H, Möllers T, Perna L, Brenner H (2018) The genetic risk of Alzheimer's disease beyond APOE $\epsilon 4$: systematic review of Alzheimer's genetic risk scores. *Transl Psychiatry* 8:1–9. <https://doi.org/10.1038/s41398-018-0221-8>
39. Lane CA, Barnes J, Nicholas JM, Sudre CH, Cash DM, Parker TD, Malone IB, Lu K, James SN, Keshavan A (2019) Associations between blood pressure across adulthood and late-life brain structure and pathology in the neuroscience substudy of the 1946 British birth cohort (Insight 46): an epidemiological study. *Lancet Neurol* 18:942–952
40. McGrath ER, Beiser AS, DeCarli C, Plourde KL, Vasan RS, Greenberg SM, Seshadri S (2017) Blood pressure from mid- to late life and risk of incident dementia. *Neurology* 89: 2447–2454. <https://doi.org/10.1212/WNL.0000000000004741>
41. Emmerzaal TL, Kiliaan AJ, Gustafson DR (2015) 2003–2013: a decade of body mass index, Alzheimer's disease, and dementia. *J Alzheimer's Dis* 43:739–755. <https://doi.org/10.3233/JAD-141086>
42. Morris JC (1993) The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* 43:2412–2414. <http://www.ncbi.nlm.nih.gov/pubmed/8232972>
43. Folstein MF, Folstein SE, McHugh PR (1975) Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *J Psychiatric Research* 12:189–198. [https://doi.org/10.1016/0022-3956\(75\)90026-6](https://doi.org/10.1016/0022-3956(75)90026-6)
44. Rosen WG, Mohs RC, Davis KL (1984) A new rating scale for Alzheimer's disease. *Am J Psychiatry* 141. <https://doi.org/10.1176/AJP.141.11.1356>
45. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings JL, Chertkow H (2005) The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* 53:695–699. <https://doi.org/10.1111/J.1532-5415.2005.53221.X>
46. Barnett JH, Blackwell AD, Sahakian BJ, Robbins TW (2016) The Paired Associates Learning (PAL) test: 30 years of CANTAB translational neuroscience from laboratory to bedside in dementia research. *Curr Top Behav Neurosci* 28:449–474. https://doi.org/10.1007/7854_2015_5001
47. Brooker H, Williams G, Hampshire A, Corbett A, Aarsland D, Cummings J, Molinuevo JL, Atri A, Ismail Z, Creese B, Fladby T, Thim-Hansen C, Wesnes K, Ballard C (2020) Flame: a computerized neuropsychological composite for trials in early dementia. *Alzheimer's Dementia* 12. <https://doi.org/10.1002/DAD2.12098>
48. Blennow K, Zetterberg H (2018) Biomarkers for Alzheimer's disease: current status and prospects for the future. *J Intern Med* 284: 643–663. <https://doi.org/10.1111/JOIM.12816>
49. Fourier A, Portelius E, Zetterberg H, Blennow K, Quadrio I, Perret-Liaudet A (2015) Pre-analytical and analytical factors influencing Alzheimer's disease cerebrospinal fluid biomarker variability. *Clin Chim Acta* 449:9–15. <https://doi.org/10.1016/J.CCA.2015.05.024>
50. Mattsson N, Andreasson U, Persson S, Carrillo MC, Collins S, Chalbot S, Cutler N, Dufour-Rainfray D, Fagan AM, Heegaard NH, Hsiung GYR, Hyman B, Iqbal K,

- Lachno DR, Lleó A, Lewczuk P, Molinuevo JL, et al (2013) CSF biomarker variability in the Alzheimer's association quality control program. *Alzheimer's Dementia* 9:251–261. <https://doi.org/10.1016/J.JALZ.2013.01.010>
51. Nakamura A, Kaneko N, Villemagne VL, Kato T, Doecke J, Doré V, Fowler C, Li QX, Martins R, Rowe C, Tomita T, Matsuzaki K, Ishii K, Ishii K, Arahata Y, Iwamoto S, Ito K, Tanaka K, Masters CL, Yanagisawa K (2018) High performance plasma amyloid- β biomarkers for Alzheimer's disease. *Nature* 554:249–254. <https://doi.org/10.1038/nature25456>
52. Janelidze S, Bali D, Ashton NJ, Barthélémy NR, Vanbrabant J, Stoops E, Vanmechelen E, He Y, Dolado AO, Triana-Baltzer G, Pontecorvo MJ, Zetterberg H, Kolb H, Vandijck M, Blennow K, Bateman RJ, Hansson O (2022) Head-to-head comparison of 10 plasma phospho-tau assays in prodromal Alzheimer's disease. *Brain* <https://doi.org/10.1093/BRAIN/AWAC333>
53. Milà-Alomà M, Ashton NJ, Shekari M, Salvadó G, Ortiz-Romero P, Montoliu-Gaya L, Benedet AL, Karikari TK, Lantero-Rodriguez J, Vanmechelen E, Day TA, González-Escalante A, Sánchez-Benavides G, Mingüillon C, Fauria K, Molinuevo JL, Dage JL, Zetterberg H, Gispert JD, Suárez-Calvet M, Blennow K (2022) Plasma p-tau231 and p-tau217 as state markers of amyloid- β pathology in preclinical Alzheimer's disease. *Nat Med* 28:1797–1801. <https://doi.org/10.1038/s41591-022-01925-w>
54. Villemagne VL, Burnham S, Bourgeat P, Brown B, Ellis KA, Salvado O, Szoéke C, Macaulay SL, Martins R, Maruff P, Ames D, Rowe CC, Masters CL (2013) Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. *Lancet Neurol* 12:357–367. [https://doi.org/10.1016/S1474-4422\(13\)70044-9](https://doi.org/10.1016/S1474-4422(13)70044-9)
55. Gordon BA, Blazey TM, Su Y, Hari-Raj A, Dincer A, Flores S, Christensen J, McDade E, Wang G, Xiong C, Cairns NJ, Hassenstab J, Marcus DS, Fagan AM, Clifford RJ Jr, Hornbeck RC, Paumier KL, Ances BM, Berman SB, Brickman AM, Cash DM, Chhatwal JP, Correia S, Förster S, Fox NC, Graff-Radford NR, la Fougère C, Levin J, Masters CL, Rossor MN, Salloway S, Saykin AJ, Schofield PR, Thompson PM, Weiner MM, Holtzman DM, Raichle ME, Morris JC, Bateman RJ, Benzinger TLS (2018) Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer's disease: a longitudinal study. *Lancet Neurol* 17:241–250. [https://doi.org/10.1016/S1474-4422\(18\)30028-0](https://doi.org/10.1016/S1474-4422(18)30028-0)
56. Aizenstein HJ, Nebes RD, Saxton JA, Price JC, Mathis CA, Tsopelas ND, Ziolkowski SK, James JA, Snitz BE, Houck PR, Bi W, Cohen AD, Lopresti BJ, DeKosky ST, Halligan EM, Klunk WE (2008) Frequent amyloid deposition without significant cognitive impairment among the elderly. *Arch Neurol* 65:1509. <https://doi.org/10.1001/ARCHNEUR.65.11.1509>
57. Rowe CC, Bourgeat P, Ellis KA, Brown B, Lim YY, Mulligan R, Jones G, Maruff P, Woodward M, Price R, Robins P, Tochon-Danguy H, O'Keefe G, Pike KE, Yates P, Szoéke C, Salvado O, Macaulay SL, O'Meara T, Head R, Cobián L, Savage G, Martins R, Masters CL, Ames D, Villemagne VL (2013) Predicting Alzheimer disease with β -amyloid imaging: results from the Australian imaging, biomarkers, and lifestyle study of ageing. *Ann Neurol* 74:905–913. <https://doi.org/10.1002/ANA.24040>
58. Johnson KA, Sperling RA, Gidicsin CM, Carmasin JS, Maye JE, Coleman RE, Reiman EM, Sabbagh MN, Sadowsky CH, Fleisher AS, Doraiswamy PM, Carpenter AP, Clark CM, Joshi AD, Lu M, Grundman M, Mintun MA, Pontecorvo MJ, Skovronsky DM (2013) Florbetapir (fl-18-av-45) PET to assess amyloid burden in Alzheimer's disease dementia, mild cognitive impairment, and normal aging. *Alzheimer's Dementia* 9. <https://doi.org/10.1016/J.JALZ.2012.10.007>
59. Hatashita S, Yamasaki H, Suzuki Y, Tanaka K, Wakebe D, Hayakawa H (2014) [18f]flutemetamol amyloid-beta PET imaging compared with [11c]PIB across the spectrum of Alzheimer's disease. *Euro J Nucl Med Mol Imaging* 41:290–300. <https://doi.org/10.1007/S00259-013-2564-Y>
60. Leuzy A, Chiotis K, Lemoine L, Gillberg PG, Almkvist O, Rodriguez-Vieitez E, Nordberg A (2019) Tau PET imaging in neurodegenerative tauopathies—still a challenge. *Mol Psychiatry* 24:1112–1134. <https://doi.org/10.1038/S41380-018-0342-8>
61. Braak H, Braak E (1991) Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol* 82:239–259. <https://doi.org/10.1007/BF00308809>
62. Cho H, Choi JY, Lee HS, Lee JH, Ryu YH, Lee MS, Jack CR, Lyoo CH (2019) Progressive tau accumulation in Alzheimer disease: 2-year follow-up study. *J Nucl Med* 60:

- 1611–1621. <https://doi.org/10.2967/JNUMED.118.221697>
63. Schöll M, Lockhart SN, Schonhaut DR, O’Neil JP, Janabi M, Ossenkoppele R, Baker SL, Vogel JW, Faria J, Schwimmer HD, Rabinovici GD, Jagust WJ (2016) PET imaging of tau deposition in the aging human brain. *Neuron* 89:971–982. <https://doi.org/10.1016/J.NEURON.2016.01.028>
64. Jones DT, Graff-Radford J, Lowe VJ, Wiste HJ, Gunter JL, Senjem ML, Botha H, Kantarci K, Boeve BF, Knopman DS, Petersen RC, Clifford RJ Jr (2017) Tau, amyloid, and cascading network failure across the Alzheimer’s disease spectrum. *Cortex* 97:143–159. <https://doi.org/10.1016/j.cortex.2017.09.018>
65. Vogel JW, Young AL, Oxtoby NP, Smith R, Ossenkoppele R, Strandberg OT, Joie RL, Aksman LM, Grothe MJ, Iturria-Medina Y, Weiner M, Aisen P, Petersen R, Jr CRJ, Jagust W, Trojanowski JQ, Toga AW, et al (2021) Four distinct trajectories of tau deposition identified in Alzheimer’s disease. *Nat Med* 27:871–881. <https://doi.org/10.1038/s41591-021-01309-6>
66. Mattsson-Carlsgren N, Andersson E, Janelidze S, Ossenkoppele R, Insel P, Strandberg O, Zetterberg H, Rosen HJ, Rabinovici G, Chai X, Blennow K, Dage JL, Stomrud E, Smith R, Palmqvist S, Hansson O (2020) $\text{A}\beta$ deposition is associated with increases in soluble and phosphorylated tau that precede a positive tau PET in Alzheimer’s disease. *Sci Adv* 6. https://doi.org/10.1126/SCIAADV.AAZ2387/SUPPL_FILE/AAZ2387_SM.PDF
67. Gordon BA, Blazey TM, Christensen J, Dincer A, Flores S, Keefe S, Chen C, Su Y, McDade EM, Wang G, Li Y, Hassenstab J, Aschenbrenner A, Hornbeck R, Clifford RJ Jr, Ances BM, Berman SB, Brosch JR, Galasko D, Gauthier S, Lah JJ, Masellis M, van Dyck CH, Mintun MA, Klein G, Ristic S, Cairns NJ, Marcus DS, Xiong C, Holtzman DM, Raichle ME, Morris JC, Bateman RJ, Benzinger TLS (2019) Tau PET in autosomal dominant Alzheimer’s disease: relationship with cognition, dementia and other biomarkers. *Brain* 142:1063–1076. <https://doi.org/10.1093/brain/awz019>
68. Quiroz YT, Sperling RA, Norton DJ, Baena A, Arboleda-Velasquez JF, Cosio D, Schultz A, Lapoint M, Guzman-Velez E, Miller JB, Kim LA, Chen K, Tariot PN, Lopera F, Reiman EM, Johnson KA (2018) Association between amyloid and tau accumulation in young adults with autosomal dominant Alzheimer disease. *JAMA Neurol* 02129:1–9. <https://doi.org/10.1001/jamaneurol.2017.4907>
69. Ossenkoppele R, Schonhaut DR, Schöll M, Lockhart SN, Ayakta N, Baker SL, O’Neil JP, Janabi M, Lazaris A, Cantwell A, Vogel J, Santos M, Miller ZA, Bettcher BM, Vossel KA, Kramer JH, Gorno-Tempini ML, Miller BL, Jagust WJ, Rabinovici GD (2016) Tau PET patterns mirror clinical and neuroanatomical variability in Alzheimer’s disease. *Brain* 139:1551–1567. <https://doi.org/10.1093/BRAIN/AWW027>
70. Jones DT, Knopman DS, Graff-Radford J, Syrjanen JA, Senjem ML, Schwarz CG, Dhele C, Wszolek Z, Rademakers R, Kantarci K, Petersen RC, Jr CRJ, Lowe VJ, Boeve BF (2018) In vivo 18F-AV-1451 tau PET signal in MAPT mutation carriers varies by expected tau isoforms. *Neurology* 90: e947–e954. <https://doi.org/10.1212/WNL.0000000000005117>
71. Smith R, Santillo AF, Waldö ML, Strandberg O, Berron D, Vestberg S, van Westen D, van Swieten J, Honer M, Hansson O (2019) 18F-Flortaucipir in TDP-43 associated frontotemporal dementia. *Sci Rep* 9:1–10. <https://doi.org/10.1038/s41598-019-42625-9>
72. Beyer L, Brendel M (2021) Imaging of tau pathology in neurodegenerative diseases: an update. *Semin Nucl Med* 51:253–263. <https://doi.org/10.1053/J.SEMNUCLMED.2020.12.004>
73. Chan D, Fox NC, Scialli RI, Crum WR, Whitwell JL, Leschziner G, Rossor AM, Stevens JM, Cipolotti L, Rossor MN (2001) Patterns of temporal lobe atrophy in semantic dementia and Alzheimer’s disease. *Ann Neurol* 49:433–442. <https://doi.org/10.1002/ANA.92>
74. Seok ML, Katsifis A, Villemagne VL, Best R, Jones G, Saling M, Bradshaw J, Merory J, Woodward M, Hopwood M, Rowe CC (2009) The 18F-FDG PET cingulate island sign and comparison to 123I- β -CIT SPECT for diagnosis of dementia with Lewy bodies. *J Nucl Med* 50:1638–1645. <https://doi.org/10.2967/JNUMED.109.065870>
75. Vemuri P, Simon G, Kantarci K, Whitwell JL, Senjem ML, Przybelski SA, Gunter JL, Josephs KA, Knopman DS, Boeve BF, Ferman TJ, Dickson DW, Parisi JE, Petersen RC, Jack CR (2011) Antemortem differential diagnosis of dementia pathology using structural MRI: differential-STAND. *Neuroimage* 55:522–531. <https://doi.org/10.1016/J.NEUROIMAGE.2010.12.073>

76. Wardlaw JM, Smith EE, Biessels GJ, Cordonnier C, Fazekas F, Frayne R, Lindley RI, O'Brien JT, Barkhof F, Benavente OR, Black SE, Brayne C, Breteler M, Chabriat H, DeCarli C, de Leeuw FE, Doubal F, Duering M, Fox NC, Greenberg S, Hachinski V, Kilimann I, Mok V, van Oostenbrugge R, Pantoni L, Speck O, Stephan BC, Teipel S, Viswanathan A, Werring D, Chen C, Smith C, van Buchem M, Norrvig B, Gorelick PB, Dichgans M (2013) Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol 12:822–838. [https://doi.org/10.1016/S1474-4422\(13\)70124-8](https://doi.org/10.1016/S1474-4422(13)70124-8)
77. Fiford CM, Manning EN, Bartlett JW, Cash DM, Malone IB, Ridgway GR, Lehmann M, Leung KK, Sudre CH, Ourselin S, Biessels GJ, Carmichael OT, Fox NC, Cardoso MJ, Barnes J (2017) White matter hyperintensities are associated with disproportionate progressive hippocampal atrophy. Hippocampus 27: 249–262. <https://doi.org/10.1002/HIPO.22690>
78. Rabin JS, Schultz AP, Hedden T, Viswanathan A, Marshall GA, Kilpatrick E, Klein H, Buckley RF, Yang HS, Properzi M, Rao V, Kirn DR, Papp KV, Rentz DM, Johnson KA, Sperling RA, Chhatwal JP (2018) Interactive associations of vascular risk and β -amyloid burden with cognitive decline in clinically normal elderly individuals: findings from the Harvard Aging Brain Study. JAMA Neurol. <https://doi.org/10.1001/JAMANEUROL.2018.1123>
79. Sexton CE, Kalu UG, Filippini N, Mackay CE, Ebmeier KP (2011) A meta-analysis of diffusion tensor imaging in mild cognitive impairment and Alzheimer's disease. Neurobiol Aging 32:2322.e5–2322.e18. <https://doi.org/10.1016/J.NEUROBIOLAGING.2010.05.019>
80. Mito R, Raffelt D, Dhollander T, Vaughan DN, Tournier JD, Salvado O, Brodtmann A, Rowe CC, Villemagne VL, Connelly A (2018) Fibre-specific white matter reductions in Alzheimer's disease and mild cognitive impairment. Brain. <https://doi.org/10.1093/brain/awx355>
81. Slattery CF, Zhang J, Paterson RW, Foulkes AJ, Carton A, Macpherson K, Mancini L, Thomas DL, Modat M, Toussaint N, Cash DM, Thornton JS, Henley SM, Crutch SJ, Alexander DC, Ourselin S, Fox NC, Zhang H, Schott JM (2017) ApoE influences regional white-matter axonal density loss in Alzheimer's disease. Neurobiol Aging 57:8–17. <https://doi.org/10.1016/j.neurobiolaging.2017.04.021>
82. Weston PSJ, Simpson IJA, Ryan NS, Ourselin S, Fox NC (2015) Diffusion imaging changes in grey matter in Alzheimer's disease: a potential marker of early neurodegeneration. Alzheimer's Res Therapy 7:47. <https://doi.org/10.1186/s13195-015-0132-3>
83. Contreras JA, Avena-Koenigsberger A, Risacher SL, West JD, Tallman E, McDonald BC, Farlow MR, Apostolova LG, Goñi J, Dzemidzic M, Wu YC, Kessler D, Jeub L, Fortunato S, Saykin AJ, Sporns O (2019) Resting state network modularity along the prodromal late onset Alzheimer's disease continuum. NeuroImage Clin 22:101687. <https://doi.org/10.1016/J.NICL.2019.101687>
84. Damoiseaux JS, Prater KE, Miller BL, Greicius MD (2012) Functional connectivity tracks clinical deterioration in Alzheimer's disease. Neurobiol Aging 33:828.e19–828.e30. <https://doi.org/10.1016/J.NEUROBIOLAGING.2011.06.024>
85. Mormino EC, Smiljic A, Hayenga AO, Onami SH, Greicius MD, Rabinovici GD, Janabi M, Baker SL, Yen IV, Madison CM, Miller BL, Jagust WJ (2011) Relationships between beta-amyloid and functional connectivity in different components of the default mode network in aging. Cereb Cortex 21: 2399–2407. <https://doi.org/10.1093/CERCOR/BHR025>
86. Leng K, Li E, Eser R, Piergies A, Sit R, Tan M, Neff N, Li SH, Rodriguez RD, Suemoto CK, Leite REP, Ehrenberg AJ, Pasqualucci CA, Seeley WW, Spina S, Heinsen H, Grinberg LT, Kampmann M (2021) Molecular characterization of selectively vulnerable neurons in Alzheimer's disease. Nat Neurosci 24:276–287. <https://doi.org/10.1038/s41593-020-00764-7>
87. Crist AM, Hinkle KM, Wang X, Moloney CM, Matchett BJ, Labuzan SA, Frankenhauser I, Azu NO, Liesinger AM, Lesser ER, Serie DJ, Quicksall ZS, Patel TA, Carnwath TP, DeTure M, Tang X, Petersen RC, Duara R, Graff-Radford NR, Allen M, Carrasquillo MM, Li H, Ross OA, Ertekin-Taner N, Dickson DW, Asmann YW, Carter RE, Murray ME (2021) Transcriptomic analysis to identify genes associated with selective hippocampal vulnerability in Alzheimer's disease. Nat Commun 12:1–17. <https://doi.org/10.1038/s41467-021-22399-3>

88. Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, Harvey D, Clifford RJ Jr, Jagust W, Liu E, Morris JC, Petersen RC, Saykin AJ, Schmidt ME, Shaw L, Siuciak JA, Soares H, Toga AW, Trojanowski JQ (2011) The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's & Dementia* 1–67. <https://doi.org/10.1016/j.jalz.2011.09.172>
89. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM (2012) FSL. *Neuroimage* 62(2):782–790
90. Ashburner J (2012) SPM: a history. *Neuroimage* 62(2):791–800
91. Ashburner J, Friston KJ (2005) Unified segmentation. *Neuroimage* 26(3):839–851
92. Zhang Y, Brady M, Smith S (2001) Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging* 20(1):45–57
93. Kumar P, Nagar P, Arora C, Gupta A (2018) U-segnet: fully convolutional neural network based automated brain tissue segmentation tool. Proceedings—international conference on image processing, ICIP, pp 3503–3507. <https://doi.org/10.1109/ICIP.2018.8451295>
94. Shen D, Davatzikos C (2002) HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging* 21(11):1421–1439
95. Fischl B (2012) Freesurfer. *Neuroimage* 62: 774–81. <https://doi.org/10.1016/j.neuroimage.2012.01.021>
96. Cardoso MJ, Modat M, Wolz R, Melbourne A, Cash DM, Rueckert D, Ourselin S (2015) Geodesic information flows: spatially-variant graphs and their application to segmentation and fusion. *IEEE Trans Med Imaging* 34:1976–1988. <https://doi.org/10.1109/TMI.2015.2418298>
97. Brébisson AD, Montana G (2015) Deep neural networks for anatomical brain segmentation. *IEEE Comput Soc Conf Comput Vis Pattern Recogn Workshops 2015-October*:20–28. <https://doi.org/10.1109/CVPRW.2015.7301312>
98. Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M (2020) FastSurfer-a fast and accurate deep learning based neuroimaging pipeline. *Neuroimage* 219:117012
99. Billot B, Greve DN, Puonti O, Thielscher A, Van Leemput K, Fischl B, Dalca AV, Iglesias JE (2021) SynthSeg: domain randomisation for segmentation of brain MRI scans of any contrast and resolution. arXiv preprint arXiv:210709559
100. Coupé P, Manjón JV, Fonov V, Pruessner J, Robles M, Collins DL (2011) Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. *Neuroimage* 54(2):940–954
101. Collins DL, Pruessner JC (2010) Towards accurate, automatic segmentation of the hippocampus and amygdala from mri by augmenting animal with a template library and label fusion. *Neuroimage* 52(4):1355–1366
102. Chupin M, Gérardin E, Cuignet R, Boutet C, Lemieux L, Lehéricy S, Benali H, Garneron L, Colliot O (2009) Fully automatic hippocampus segmentation and classification in alzheimer's disease and mild cognitive impairment applied on data from adni. *Hippocampus* 19(6):579–587
103. Tong T, Wolz R, Coupé P, Hajnal JV, Rueckert D, Initiative ADN et al (2013) Segmentation of mr images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. *Neuroimage* 76:11–23
104. Xie L, Wisse LE, Pluta J, de Flores R, Piskin V, Manjón JV, Wang H, Das SR, Ding SL, Wolk DA et al (2019) Automated segmentation of medial temporal lobe subregions on in vivo t1-weighted mri in early stages of alzheimer's disease. *Hum Brain Mapp* 40(12): 3431–3451
105. Pipitone J, Park MTM, Winterburn J, Lett TA, Lerch JP, Pruessner JC, Lepage M, Voineskos AN, Chakravarty MM, Initiative ADN, et al (2014) Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. *Neuroimage* 101:494–512
106. Yushkevich PA, Wang H, Pluta J, Das SR, Craige C, Avants BB, Weiner MW, Mueller S (2010) Nearly automatic segmentation of hippocampal subfields in in vivo focal t2-weighted MRI. *Neuroimage* 53(4): 1208–1224
107. Manjón JV, Romero JE, Coupe P (2022) A novel deep learning based hippocampus sub-field segmentation method. *Sci Rep* 12:1–9. <https://doi.org/10.1038/s41598-022-05287-8>
108. Vanderbecq Q, Xu E, Ströer S, Couvy-Duchesne B, Melo MD, Dormont D, Colliot O (2020) Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients. *NeuroImage: Clin* 27:102357

109. Sudre C, Cardoso MJ, Bouvy W, Biessels G, Barnes J, Ourselin S (2015) Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. *IEEE Trans Med Imaging* 34:1–1. <https://doi.org/10.1109/TMI.2015.2419072>
110. Boutinaud P, Tsuchida A, Laurent A, Adonias F, Hanifehlou Z, Nozais V, Verrecchia V, Lampe L, Zhang J, Zhu YC, Tzourio C, Mazoyer B, Joliot M (2021) 3D segmentation of perivascular spaces on T1-weighted 3 Tesla MR images with a convolutional autoencoder and a U-shaped neural network. *Front Neuroinform* 15. <https://doi.org/10.3389/FNINF.2021.641600>
111. Wu J, Zhang Y, Wang K, Tang X (2019) Skip connection U-Net for white matter hyperintensities segmentation from MRI. *IEEE Access* 7:155194–155202. <https://doi.org/10.1109/ACCESS.2019.2948476>
112. Klöppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, Fox NC, Jack J Clifford R, Ashburner J, Frackowiak RSJ (2008) Automatic classification of MR scans in Alzheimer's disease. *Brain* 131(3):681–689. <https://doi.org/10.1093/brain/awm319>
113. Vemuri P, Gunter JL, Senjem ML, Whitwell JL, Kantarci K, Knopman DS, Boeve BF, Petersen RC, Jack Jr CR (2008) Alzheimer's disease diagnosis in individual subjects using structural mr images: validation studies. *NeuroImage* 39(3):1186–1197
114. Gerardin E, Chételat G, Chupin M, Cuingnet R, Desgranges B, Kim HS, Niethammer M, Dubois B, Lehéricy S, Garnerot L, et al (2009) Multidimensional classification of hippocampal shape features discriminates alzheimer's disease and mild cognitive impairment from normal aging. *Neuroimage* 47(4):1476–1486
115. Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert MO, Chupin M, Benali H, Colliot O, Initiative ADN, et al (2011) Automatic classification of patients with alzheimer's disease from structural MRI: a comparison of ten methods using the adni database. *Neuroimage* 56(2):766–781
116. Coupé P, Eskildsen SF, Manjón JV, Fonov VS, Collins DL (2012) Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. *Neuroimage* 59:3736–3747. <https://doi.org/10.1016/J.NEUROIMAGE.2011.10.080>
117. Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D (2013) Random forest-based similarity measures for multimodal classification of Alzheimer's disease. *Neuroimage* 65:167–175. <https://doi.org/https://doi.org/10.1016/j.neuroimage.2012.09.065>
118. Tong T, Wolz R, Gao Q, Guerrero R, Hajnal JV, Rueckert D, Initiative ADN et al (2014) Multiple instance learning for classification of dementia in brain MRI. *Med Image Anal* 18(5):808–818
119. Lian C, Liu M, Zhang J, Shen D (2018) Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE Trans Pattern Anal Mach Intell* 42(4):880–893
120. Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ (2011) Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol Aging* 32(12):2322.e19–2322.e27. <https://doi.org/https://doi.org/10.1016/j.neurobiolaging.2010.05.023>
121. Young J, Modat M, Cardoso MJ, Mendelson A, Cash D, Ourselin S, Initiative ADN, et al (2013) Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage Clin* 2:735–745
122. Coupé P, Fonov VS, Bernard C, Zandifar A, Eskildsen SF, Helmer C, Manjón JV, Amieva H, Dartigues JF, Allard M, et al (2015) Detection of alzheimer's disease signature in mr images seven years before conversion to dementia: toward an early individual prognosis. *Hum Brain Mapp* 36(12):4758–4770
123. Jie B, Zhang D, Cheng B, Shen D, Initiative ADN (2015) Manifold regularized multitask feature learning for multimodality disease classification. *Hum Brain Mapp* 36(2):489–507
124. Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J, Initiative ADN et al (2015) Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage* 104:398–412
125. Rathore S, Habes M, Iftikhar MA, Shacklett A, Davatzikos C (2017) A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *Neuroimage* 155:530–548
126. Jo T, Nho K, Saykin AJ (2019) Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using

- neuroimaging data. *Front Aging Neurosci* 11:220. <https://doi.org/10.3389/fnagi.2019.00220>
127. Ebrahimpournavieh MA, Luo S, Chiong R (2020) Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review. *Comput Methods Program Biomed* 187:105242. <https://doi.org/https://doi.org/10.1016/j.cmpb.2019.105242>
 128. Tanveer M, Richhariya B, Khan RU, Rashid AH, Khanna P, Prasad M, Lin CT (2020) Machine learning techniques for the diagnosis of Alzheimer's disease: a review. *ACM Trans Multimedia Comput Commun Appl* 16(1s). <https://doi.org/10.1145/3344998>
 129. Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Dormont D, Durrleman S, Burgos N, Colliot O (2020) Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. *Med Image Anal* 63:101694
 130. Ansart M, Epelbaum S, Bassignana G, Bône A, Bottani S, Cattai T, Couronné R, Faouzi J, Koval I, Louis M, Thibeau-Sutre E, Wen J, Wild A, Burgos N, Dormont D, Colliot O, Durrleman S (2021) Predicting the progression of mild cognitive impairment using machine learning: a systematic, quantitative and critical review. *Med Image Anal* 67: 101848. <https://doi.org/https://doi.org/10.1016/j.media.2020.101848>
 131. Canas LS, Sudre CH, De Vita E, Nihat A, Mok TH, Slattery CF, Paterson RW, Foulkes AJ, Hyare H, Cardoso MJ, Thornton J, Schott JM, Barkhof F, Collinge J, Ourselin S, Mead S, Modat M (2019) Prion disease diagnosis using subject-specific imaging biomarkers within a multi-kernel Gaussian process. *NeuroImage Clin* 24:102051. <https://doi.org/https://doi.org/10.1016/j.nicl.2019.102051>
 132. Marinescu RV, Bron EE, Oxtoby NP, Young AL, Toga AW, Weiner MW, Barkhof F, Fox NC, Golland P, Klein S, et al (2020) Predicting Alzheimer's disease progression: results from the TADPOLE Challenge. *Alzheimer's & Dementia* 16:e039538
 133. Samper-Gonzalez J, Burgos N, Bottani S, Habert MO, Evgeniou T, Epelbaum S, Colliot O (2019) Reproducible evaluation of methods for predicting progression to Alzheimer's disease from clinical and neuroimaging data. In: Medical imaging 2019: image processing, vol 10949, pp 221–233
 134. Wang H, Nie F, Huang H, Risacher SL, Saykin AJ, Shen L (2012) Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics* 28(12):i127–i136
 135. Peng J, An L, Zhu X, Jin Y, Shen D (2016) Structured sparse kernel learning for imaging genetics based Alzheimer's disease diagnosis. In: International conference on medical image computing and computer-assisted intervention MICCAI, pp 70–78
 136. Samper-González J, Burgos N, Bottani S, Fontanella S, Lu P, Marcoux A, Routier A, Guillou J, Bacci M, Wen J et al (2018) Reproducible evaluation of classification methods in Alzheimer's disease: framework and application to MRI and PET data. *Neuroimage* 183: 504–521
 137. Wyman BT, Harvey DJ, Crawford K, a Bernstein M, Carmichael O, Cole PE, Crane PK, Decarli C, Fox NC, Gunter JL, Hill D, Killiany RJ, Pachai C, Schwarz AJ, Schuff N, Senjem ML, Suhy J, Thompson PM, Weiner M, Clifford RJ Jr (2012) Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's Dementia* 1–6. <https://doi.org/10.1016/j.jalz.2012.06.004>
 138. Bron EE, Smits M, Van Der Flier WM, Vrenken H, Barkhof F, Scheltens P, Papma JM, Steketee RM, Orellana CM, Meijboom R, et al (2015) Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDEmentia challenge. *Neuroimage* 111:562–579
 139. Davatzikos C, Resnick SM, Wu X, Parmpi P, Clark CM (2008) Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *Neuroimage* 41(4): 1220–1227
 140. Koikkalainen J, Rhodius-Meester H, Tolonen A, Barkhof F, Tijms B, Lemstra AW, Tong T, Guerrero R, Schuh A, Ledig C, et al (2016) Differential diagnosis of neurodegenerative diseases using structural MRI data. *NeuroImage Clin* 11:435–449
 141. Morin A, Samper-Gonzalez J, Bertrand A, Ströer S, Dormont D, Mendes A, Coupé P, Ahidjan J, Lévy M, Samri D et al (2020) Accuracy of MRI classification algorithms in a tertiary memory center clinical routine cohort. *J Alzheimer's Dis* 74(4):1157–1166
 142. Jack Jr CR, Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, Shaw LM, Vemuri P, Wiste HJ, Weigand SD, et al (2013) Tracking pathophysiological processes in Alzheimer's disease: an updated

- hypothetical model of dynamic biomarkers. *Lancet Neurol* 12(2):207–216
143. Fonteijn HM, Modat M, Clarkson MJ, Barnes J, Lehmann M, Hobbs NZ, Scialli RI, Tabrizi SJ, Ourselin S, Fox NC, Alexander DC (2012) An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *Neuroimage* 60(3):1880–1889. <https://doi.org/https://doi.org/10.1016/j.neuroimage.2012.01.062>
144. Oxtoby NP, Young AL, Cash DM, Benzinger TL, Fagan AM, Morris JC, Bateman RJ, Fox NC, Schott JM, Alexander DC (2018) Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain* 141(5): 1529–1544
145. Young AL, Oxtoby NP, Daga P, Cash DM, Fox NC, Ourselin S, Schott JM, Alexander DC (2014) A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* 137(9):2564–2577
146. Venkatraghavan V, Bron EE, Niessen WJ, Klein S, et al (2019) Disease progression timeline estimation for alzheimer's disease using discriminative event based modeling. *NeuroImage* 186:518–532
147. Firth NC, Primativo S, Brotherhood E, Young AL, Yong KX, Crutch SJ, Alexander DC, Oxtoby NP (2020) Sequences of cognitive decline in typical Alzheimer's disease and posterior cortical atrophy estimated using a novel event-based model of disease progression. *Alzheimer's Dementia* 16(7):965–973
148. Panman JL, Venkatraghavan V, Van Der Ende EL, Steketee RM, Jiskoot LC, Poos JM, Doppler EG, Meeter LH, Kaat LD, Rombouts SA, et al (2021) Modelling the cascade of biomarker changes in GRN-related frontotemporal dementia. *J Neurol Neurosurg Psychiatry* 92(5):494–501
149. Montal V, Vilaplana E, Alcolea D, Pegueroles J, Pasternak O, González-Ortiz S, Clarimón J, Carmona-Iragui M, Illán-Gala I, Morenas-Rodríguez E, Ribosa-Nogué R, Sala I, Sánchez-Saudinós MB, García-Sebastian M, Villanúa J, Izagirre A, Estanga A, Ecay-Torres M, Iriondo A, Clerigue M, Tainta M, Pozueta A, González A, Martínez-Heras E, Llufriu S, Blesa R, Sanchez-Juan P, Martínez-Lage P, Lleó A, Fortea J (2018) Cortical microstructural changes along the alzheimer's disease continuum. *Alzheimer's Dementia* 14:340–351. <https://doi.org/10.1016/j.jalz.2017.09.013>
150. Jedynak BM, Lang A, Liu B, Katz E, Zhang Y, Wyman BT, Raunig D, Jedynak CP, Caffo B, Prince JL (2012) A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease neuroimaging initiative cohort. *Neuroimage* 63(3):1478–1486. <https://doi.org/https://doi.org/10.1016/j.neuroimage.2012.07.059>
151. Schiratti JB, Allissonnière S, Colliot O, Durrellman S (2017) A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. *J Mach Learn Res* 18(1):4840–4872
152. Koval I, Bône A, Louis M, Lartigue T, Bottani S, Marcoux A, Samper-Gonzalez J, Burgos N, Charlier B, Bertrand A, et al (2021) AD Course Map charts Alzheimer's disease progression. *Sci Rep* 11(1):1–16
153. Wijeratne PA, Alexander DC (2020) Learning transition times in event sequences: the event-based hidden markov model of disease progression. <https://doi.org/10.48550/ARXIV.2011.01023>
154. Abi Nader C, Ayache N, Frisoni GB, Robert P, Lorenzi M, for the Alzheimer's Disease Neuroimaging Initiative (2021) Simulating the outcome of amyloid treatments in Alzheimer's disease from imaging and clinical data. *Brain Commun* 3(2). <https://doi.org/10.1093/braincomms/fcab091>
155. Erus G, Doshi J, An Y, Verganelakis D, Resnick SM, Davatzikos C (2018) Longitudinally and inter-site consistent multi-atlas based parcellation of brain anatomy using harmonized atlases. *Neuroimage* 166:71. <https://doi.org/10.1016/j.neuroimage.2017.10.026>
156. Jog A, Carass A, Roy S, Pham DL, Prince JL (2015) MR image synthesis by contrast learning on neighborhood ensembles. *Med Image Anal* 24:63–76. <https://doi.org/10.1016/j.media.2015.05.002>
157. Fortin JP, Sweeney EM, Muschelli J, Crainiceanu CM, Shinohara RT (2016) Removing inter-subject technical variability in magnetic resonance imaging studies. *Neuroimage* 132: 198–212. <https://doi.org/https://doi.org/10.1016/j.neuroimage.2016.02.036>
158. Fortin JP, Parker D, Tunc B, Watanabe T, Elliott MA, Ruparel K, Roalf DR, Satterthwaite TD, Gur R, Schultz RT, Verma R, Shinohara RT (2017) Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* 161:149–170. <https://doi.org/https://doi.org/10.1101/116541>

159. Johnson WE, Li C, Rabinovic A (2006) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1):118–127. <https://doi.org/10.1093/biostatistics/kxj037>
160. Fortin JP, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, Adams P, Cooper C, Fava M, McGrath PJ, McInnis M, Phillips ML, Trivedi MH, Weissman MM, Shinohara RT (2018) Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* 167:104–120. <https://doi.org/10.1016/J.NEUROIMAGE.2017.11.024>
161. Torbati ME, Minhas DS, Ahmad G, O’Connor EE, Muschelli J, Laymon CM, Yang Z, Cohen AD, Aizenstein HJ, Klunk WE, Christian BT, Hwang SJ, Crainiceanu CM, Tudorascu DL (2021) A multi-scanner neuroimaging data harmonization using RAVEL and ComBat. *Neuroimage* 245: <https://doi.org/10.1016/J.NEUROIMAGE.2021.118703>
162. Beer JC, Tustison NJ, Cook PA, Davatzikos C, Sheline YI, Shinohara RT, Linn KA (2020) Longitudinal ComBat: a method for harmonizing longitudinal multi-scanner imaging data. *Neuroimage* 220: 117129. <https://doi.org/10.1016/j.neuroimage.2020.117129>
163. Chen AA, Beer JC, Tustison NJ, Cook PA, Shinohara RT, Shou H (2022) Mitigating site effects in covariance for machine learning in neuroimaging data. *Hum Brain Mapp* 43: 1179–1195. <https://doi.org/10.1002/HBM.25688>
164. Pomponio R, Erus G, Habes M, Doshi J, Srinivasan D, Mamourian E, Bashyam V, Nasrallah IM, Satterthwaite TD, Fan Y, Launer LJ, Masters CL, Maruff P, Zhuo C, Völzke H, Johnson SC, Fripp J, Koutsouleris N, Wolf DH, Gur R, Gur R, Morris J, Albert MS, Grabe HJ, Resnick SM, Bryan RN, Wolk DA, Shinohara RT, Shou H, Davatzikos C (2020) Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *Neuroimage* 208: 116450. <https://doi.org/10.1016/J.NEUROIMAGE.2019.116450>
165. Prado P, Birba A, Cruzat J, Santamaría-García H, Parra M, Moguilner S, Tagliazucchi E, Ibáñez A (2022) Dementia ConnEEGtome: towards multicentric harmonization of EEG connectivity in neurodegeneration. *Int J Psychophysiol* 172:24–38. <https://doi.org/10.1016/J.IJPYSCHO.2021.12.008>
166. Costa A, Bak T, Caffarra P, Caltagirone C, Ceccaldi M, Collette F, Crutch S, Sala SD, Démonet JF, Dubois B, Duzel E, Nestor P, Papageorgiou SG, Salmon E, Sikkes S, Tiraboschi P, Flier WMVD, Visser PJ, Cappa SF (2017) The need for harmonisation and innovation of neuropsychological assessment in neurodegenerative dementias in Europe: consensus document of the Joint Program for Neurodegenerative Diseases Working Group. *Alzheimer’s Res Therapy* 9:1–15. <https://doi.org/10.1186/S13195-017-0254-X>
167. Brodaty H, Woolf C, Andersen S, Barzilai N, Brayne C, Cheung KSL, Corrada MM, Crawford JD, Daly C, Gondo Y, Hagberg B, Hirose N, Holstege H, Kawas C, Kaye J, Kochan NA, Lau BHP, Lucca U, Marcon G, Martin P, Poon LW, Richmond R, Robine JM, Skoog I, Slavin MJ, Szewieczek J, Tettamanti M, Vi˜a J, Perls T, Sachdev PS (2016) ICC-dementia (International Centenarian Consortium—dementia): an international consortium to determine the prevalence and incidence of dementia in centenarians across diverse ethnoracial and socio-cultural groups. *BMC Neurol* 16:1–10. <https://doi.org/10.1186/S12883-016-0569-4/TABLES/2>
168. Srikrishna M, Pereira JB, Heckemann RA, Volpe G, van Westen D, Zettergren A, Kern S, Wahlund LO, Westman E, Skoog I, Schöll M (2021) Deep learning from MRI-derived labels enables automatic brain tissue classification on human brain CT. *Neuroimage* 244:118606. <https://doi.org/10.1016/J.NEUROIMAGE.2021.118606>
169. Bosco P, Redolfi A, Bocchetta M, Ferrari C, Mega A, Galluzzi S, Austin M, Chincarini A, Collins DL, Duchesne S, Maráchal B, Roche A, Sensi F, Wolz R, Alegret M, Assal F, Balasa M, Bastin C, Bougea A, Emek-Savaş DD, Engelborghs S, Grimmer T, Grosu G, Kramberger MG, Lawlor B, Stojmenovic GM, Marinescu M, Mecocci P, Molinuevo JL, Morais R, Niemantsverdriet E, Nobili F, Ntovas K, O’Dwyer S, Paraskevas GP, Pelini L, Picco A, Salmon E, Santana I, Sotolongo-Grau O, Spiru L, Stefanova E, Popovic KS, Tsolaki M, Yener GG, Zekry D, Frisoni GB (2017) The impact of automated hippocampal volumetry on diagnostic confidence in patients with suspected Alzheimer’s disease: a European Alzheimer’s disease consortium study. *Alzheimer’s Dementia* 13:1013–1023. <https://doi.org/10.1016/J.JALZ.2017.01.019>

170. Pemberton HG, Goodkin O, Prados F, Das RK, Vos SB, Moggridge J, Coath W, Gordon E, Barrett R, Schmitt A, Whiteley-Jones H, Burd C, Wattjes MP, Haller S, Verhooij MW, Harper L, Fox NC, Paterson RW, Schott JM, Bisdas S, White M, Ourselin S, Thornton JS, Yousry TA, Cardoso MJ, Barkhof F (2021) Automated quantitative MRI volumetry reports support diagnostic interpretation in dementia: a multi-rater, clinical accuracy study. *Euro Radiol* 31:5312–5323. <https://doi.org/10.1007/S00330-020-07455-8/TABLES/6>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made. The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 26

Machine Learning for Parkinson's Disease and Related Disorders

Johann Faouzi, Olivier Colliot, and Jean-Christophe Corvol

Abstract

Parkinson's disease is a complex heterogeneous neurodegenerative disorder characterized by the loss of dopamine neurons in the basal ganglia, resulting in many motor and non-motor symptoms. Although there is no cure to date, the dopamine replacement therapy can improve motor symptoms and the quality of life of the patients. The cardinal symptoms of this disorder are tremor, bradykinesia, and rigidity, referred to as parkinsonism. Other related disorders, such as dementia with Lewy bodies, multiple system atrophy, and progressive supranuclear palsy, share similar motor symptoms although they have different pathophysiology and are less responsive to the dopamine replacement therapy. Machine learning can be of great utility to better understand Parkinson's disease and related disorders and to improve patient care. Many challenges are still open, including early accurate diagnosis, differential diagnosis, better understanding of the pathologies, symptom detection and quantification, individual disease progression prediction, and personalized therapies. In this chapter, we review research works on Parkinson's disease and related disorders using machine learning.

Key words Clinical decision support, Deep learning, Disease understanding, Machine learning, Multiple system atrophy, Parkinson's disease, Parkinsonian syndromes, Parkinsonism, Precision medicine, Progressive supranuclear palsy

1 Introduction

Parkinson's disease (PD) is the second most frequent neurodegenerative after Alzheimer's disease, affecting more than six million individuals worldwide, a prevalence which is expected to double with the next 10 years [1]. It is characterized by the progressive degeneration of dopaminergic neurons in the *substantia nigra* associated with intracellular inclusions called Lewy bodies. These Lewy bodies are composed of protein aggregates enriched in α -synuclein. Age is the greatest risk factor, but both environmental and genetic risk factors have been associated with PD. For instance, exposure to pesticides is a well-recognized risk factor for PD, whereas caffeine intake and smoking have been demonstrated to be protective [2]. Although commonly sporadic, rare genetic forms

of the disease have been described. More than 20 loci and associated genes have been identified to be responsible for autosomal dominant or recessive forms of the disease, and more than 90 genetic risk factors have been associated with sporadic PD [3]. Although rare, genetic forms of the disease have brought important insights on the causes and pathological mechanisms of PD [4]. Among them, aggregation and spreading of misfolded α -synuclein, the protein enriched in Lewy bodies, is supposed to play a key role in the pathophysiology of the disease.

The loss of dopamine innervation of the basal ganglia network in the brain leads to the cardinal motor symptoms of the disease (parkinsonism): rest tremor, akinesia, and rigidity [2]. However, the spreading of the synucleinopathy (aggregation of α -synuclein protein) and neuronal loss outside the dopaminergic pathway is associated with other non-motor symptoms like anosmia, sleep disorders, dysautonomia, and progressive cognitive decline. Some of these symptoms, particularly anosmia, constipation, and sleep disorders, can precede the motor phase during a long prodromal phase [5].

There is no cure for PD. The therapeutic strategy relies on the dopamine replacement therapy by levodopa or dopamine agonists, which alleviate motor symptoms. However, the dopamine replacement therapy does not change the course of the disease, the progression being hampered by motor complications (motor fluctuations and abnormal movement called dyskinesia), related both to the progression of the neuronal loss and to pre- and post-synaptic plasticity induced by the treatment. In addition, the dopamine replacement therapy has no benefit on non-motor symptoms not related to the loss of dopaminergic neurons.

PD is the most frequent synucleinopathy. Other neurodegenerative diseases share some clinical and pathophysiological features of PD. Multiple system atrophy (MSA) is a rare disease associated with parkinsonism with low response to levodopa, early dysautonomia, and/or cerebellar symptoms [6]. The synucleinopathy affects the substantia nigra, but also the striatum and the cerebellum, and Lewy bodies are also observed in glial cells. There are two variants of MSA: the parkinsonian variant (MSA-P) characterized by parkinsonism and the cerebellar variant (MSA-C) characterized by gait ataxia with cerebellar dysarthria. Dementia with Lewy bodies (DLB), the second most common neurodegenerative dementia after Alzheimer's disease, is characterized by early cognitive decline, hallucinations, and levodopa-responsive motor symptoms [7]. However, whether DLB and PD with dementia are really two distinct entities is still a matter of debate. There are also other rare atypical parkinsonism syndromes, not related to a synucleinopathy. Progressive supranuclear palsy (PSP) is a tauopathy (aggregation of tau protein) characterized by a nonresponsive, axial predominant parkinsonism, early falls, supranuclear gaze palsy, and a frontal

syndrome [8]. The cortico-basal degeneration (CBD) is also a tauopathy with asymmetric parkinsonism with dystonia and cognitive dysfunction. Table 1 summarizes the characteristics of all these disorders.

Considering the complexity of these disorders, the lack of reliable biomarkers, and the overlapping clinical presentation at the early stage, there is a need for more advanced approaches to support differential diagnosis. In addition, the pathophysiology of these disorders results from the complex interplay of multiple mechanisms. One current challenge is to stratify patients according to specific mechanisms and predict individual progression profile in order to move toward a more personalized medicine. Machine learning consists in extracting information from data by computer programs without providing explicit rules on what to extract, in the sense that machines learn by themselves which information to extract. Given the complexity of Parkinson's disease and its related disorders, there still exist many challenges and open questions for which machine learning could help increase knowledge on these disorders, in particular diagnosis, disease understanding, and precision medicine, and create better clinical decision support systems. Table 2 summarizes the potential benefits of machine learning for Parkinson's disease and related disorders.

The rest of this chapter is organized as follows. We first present research works on the diagnosis of Parkinson's disease and the differential diagnosis between parkinsonian syndromes, including disease understanding (Subheading 2). We then focus on the detection and quantification of motor and non-motor symptoms in Parkinson's disease (Subheading 3). Disease progression in Parkinson's disease, with the prediction of individual progression trajectories, is presented in Subheading 4. We then describe research on the monitoring and adjustment of treatment in Parkinson's disease and discuss the limitations of machine learning in terms of causality (Subheading 5). Finally, we conclude on the existing literature and discuss open questions and research works (Subheading 6). Table 3 summarizes the studies described in this chapter.

2 Diagnosis

Having an automated model being able to accurately diagnose one or several diseases has not only a concrete utility in clinical routine, but interpreting the decision process of the model may also help better understand these diseases. To assist diagnosis, two different classification tasks are usually considered: (i) being able to differentiate PD patients from healthy controls (HC) and (ii) being able to differentiate several parkinsonian syndromes from each other.

Table 1
Main characteristics of Parkinson's disease and its related disorders

Disorder	Parkinson's disease	Multiple system atrophy	Progressive supranuclear palsy	Dementia with Lewy bodies	Cortico-basal degeneration
Pathophysiology	Aggregates of α-synuclein in dopaminergic neurons in brainstem	Aggregates of α-synuclein in neurons (cerebellum, pons, basal ganglia) and oligodendrocytes (glial cytoplasmic inclusions)	Aggregates of 4-repeat tau in astrocytes (tufts), oligodendrocytes (coiled bodies), and neurons (neurofibrillary tangles) in basal ganglia and brainstem	Aggregates of α-synuclein in neurons of the neocortex and brainstem	Aggregates of 4-repeat tau in astrocytes (astrocytic plaques), oligodendrocytes (coiled bodies), and neurons (neurofibrils) in neocortex and basal ganglia
Prevalence	100–300/100,000	5/100,000	5–10/100,000	50/100,000	1/100,000
Main symptoms	Levodopa-responsive parkinsonian syndrome Non-motor symptoms: Dysautonomia, anxiety, depression, sleep disorders, late cognitive dysfunction	Parkinsonian syndrome with poor response to levodopa Early and severe dysautonomia or cerebellar syndrome	Axial predominant parkinsonian syndrome nonresponsive to levodopa Early gait disturbance and falls Supranuclear gaze palsy Frontal syndrome	Parkinsonian syndrome responsive to levodopa Early dementia Early hallucinations or illusions Fluctuations in alertness and attention	Asymmetric parkinsonian syndrome nonresponsive to levodopa Apraxia and dystonia Early dementia Frontal syndrome
Symptomatic treatment	Levodopa, dopamine agonists, C-dopamine metabolism inhibitors, amantadine Deep brain stimulation for advanced cases	Levodopa Droxidopa, ephedrine, midodrine, or fludrocortisone for hypotension	Levodopa Cholinesterase inhibitors	Levodopa Botulinum toxin A for treating focal dystonia	Levodopa

Table 2
Summary of the potential benefits of machine learning for Parkinson's disease and related disorders

Disease stage	Potential benefits
Early PD diagnosis	Better clinical decision support systems Higher performance than current diagnostic criteria Better management and improved quality of life Potential preventive therapeutic strategies
Differential diagnosis	Better clinical decision support systems Higher performance than current diagnostic criteria Better management and improved quality of life
Symptom detection and quantification	More frequent, more robust assessment of symptoms with automatic analysis of sensor data Better management and improved quality of life
Disease progression	Identification of disease subtypes Prediction of future symptoms Treatment adjustment for potential prevention
Treatment adjustment	Better clinical decision support systems Personalized therapy Prevention of adverse events Better management and improved quality of life

2.1 Parkinson's Disease Diagnosis Compared to Healthy Subjects

Given the much larger prevalence of Parkinson's disease compared to the atypical parkinsonian syndromes, gathering data from PD patients and HC is naturally easier, especially easy-to-collect data from sensors compared to clinical, imaging, or genetic data.

Digital technologies including wearable sensors, smartphone applications, and smart algorithms receive a strongly increasing interest and begin to move toward medical applications, particularly in PD [9]. Two main types of sensor data are usually considered: voice data and motion data. Given that the cardinal symptoms of PD are motor, motion data is natural, but speech also involves motor muscles. Dysarthria, which is a motor speech disorder in which the muscles involved in producing speech are damaged, paralyzed, or weakened, is a symptom of PD.

2.1.1 PD Diagnosis Using Motion Data

Several types of sensors have been investigated to collect motion data depending on the movements of interest.

Wahid and colleagues [10] investigated the discrimination between PD patients and healthy controls using gait data collected during self-selected walking. They extracted spatial-temporal features, such as stride length, stance time, swing time, and step length, from the signals and investigated different strategies of data normalization using dimensionless equations and multiple regression and different machine learning algorithms such as naive Bayes (NB), k-nearest neighbors (kNN), support vector

Table 3
Summary of the studies reviewed in this chapter

Study	Subheading	Objectives/tasks	Modalities	Data sets	Number of subjects	Methods
Wahid et al. [10]	2.1.1	PD diagnosis (PD vs HC)	Motion data (gait)	Local	23 PD, 26 HC	NB, kNN, SVM, RF, KFD
Mirelman et al. [11]	2.1.1	PD diagnosis (PD vs HC), PD stage classification	Motion data (gait)	Local	100 HC, 58 PD-HYI, 190 PD-HYII, 84 PD-HYIII	RUSBoost
Kostikis et al. [12]	2.1.1	PD diagnosis (PD vs HC)	Motion data (hand movement)	Local	25 PD, 20 HC	NB, LR, SVM, AdaBoost, DT, RF
Kotsavasiliou et al. [13]	2.1.1	PD diagnosis (PD vs HC)	Motion data (hand movement)	Local	24 PD, 20 HC	NB, LR, SVM, AdaBoost
Amato et al. [14]	2.1.2	PD diagnosis (PD vs HC)	Voice data	IPV/C, local	28 PD, 22 HC; 26 PD, 18 HC	NB, kNN, SVM, RF, AdaBoost, gradient boosting, bagging ensemble
Jeancolas et al. [15]	2.1.2	PD diagnosis (PD vs HC)	Voice data	Local	121 PD, 100 HC	MFCC-GMM, X-vector
Jeancolas et al. [16]	2.1.2	PD diagnosis (PD vs HC)	Voice data	Local	117 PD, 41 iRBD, 98 HC	SVM
Quan et al. [17]	2.1.2	PD diagnosis (PD vs HC)	Voice data	Local	30 PD, 15 HC	DT, kNN, NB, SVM, MLP, LSTM
Adeli et al. [21]	2.1.3	PD diagnosis (PD vs HC)	Imaging data (MRI)	PPMI	374 PD, 169 NC	JFSS+robust LDA
Solana-Lavalle and Rosas-Romero [22]	2.1.3	PD diagnosis (PD vs HC)	Imaging data (MRI)	PPMI	114 PD, 49 HC, 234 PD; 110 HC	kNN, SVM, RF, NB, LR, MLP, BNN

Mudali et al. [23]	2.1.3	PD diagnosis (PD vs HC), differential diagnosis	Imaging data (FDG-PET)	Local	20 PD, 21 MSA, 17 PSP, 18 HC	DT
Huppertz et al. [26]	2.2	Differential diagnosis	Imaging data (MRI)	Local	204 PD, 106 PSP, 21 MSA-C, 60 MSA-P, 73 HC	SVM
Archer et al. [27]	2.2	Differential diagnosis	Imaging data (MRI)	Local	511 PD, 84 MSA, 129 PSP, 278 HC	SVM
Chougar et al. [28]	2.2	Differential diagnosis	Imaging data (MRI)	Local	63 PD, 21 PSP, 11 MSA-P, 12 MSA-C, 72 HC; 56 PD, 30 PSP, 24 MSA-P, 11 MSA-C, 22 HC	LR, SVM, RF
Shinde et al. [29]	2.2	Differential diagnosis	Imaging data (MRI)	Local	45 PD, 20 APS, 35 HC	CNN, gradient boosting
Jucaita et al. [30]	2.2	Differential diagnosis	Imaging data (PET)	Local	24 PD, 66 MSA	LDA
Khawaldeh et al. [31]	2.3	Parkinson's disease understanding (subthalamic nucleus activity)	Signal data (LFP)	Local	18 PD	NB
Poston et al. [32]	2.3	Parkinson's disease understanding (cognition)	Imaging data (fMRI)	Local	54 PD	SVM
Trezzini et al. [33]	2.3	Parkinson's disease understanding (metabolism)	CSF	Local	44 PD, 43 HC	LR
Vanneste et al. [34]	2.3	Parkinson's disease understanding (thalamocortical dysrhythmia)	Signal data (rs-EEG)	Local	31 PD, 264 HC, 153 tinnitus, 78 CP, 15 MDD	SVM

(continued)

Table 3
(continued)

Study	Subheading	Objectives/tasks	Modalities	Data sets	Number of subjects	Methods
Ahlrichs et al. [37]	3.1	Symptom detection and quantification (freezing of gait)	Motion data (gait)	Local	20 PD (8 with FOG, 12 without FOG)	SVM
Aich et al. [38]	3.1	Symptom detection and quantification (freezing of gait)	Motion data (gait)	Local	51 PD (36 with FOG, 15 without FOG)	SVM, kNN, DT, NB
Borzi et al. [39]	3.1	Symptom detection and quantification (freezing of gait)	Motion data (gait)	Local	11 PD	SVM, kNN, LDA, LR
Dvorani et al. [40]	3.1	Symptom detection and quantification (freezing of gait)	Motion data (gait)	Local	16 PD	SVM, AdaBoost
Park et al. [42]	3.2	Symptom detection and quantification (bradykinesia and tremor)	Motion data (resting tremor and finger tapping)	Local	55 PD	SVM
Kim et al. [43]	3.2	Symptom detection and quantification (tremor)	Motion data (tremor)	Local	92 PD	CNN, RF, NB, LR, DT, SVM, MLP
Eskofier et al. [44]	3.2	Symptom detection and quantification (bradykinesia)	Motion data (motor tasks)	Local	10 PD	CNN, AdaBoost, kNN, SVM
Abós et al. [46]	3.3	Symptom detection and quantification (cognition)	Imaging data (rs-fMRI)	Local	27 PD-MCI, 43 PD-CN, 38 HC	SVM

Betrouni et al. [47]	3.3	Symptom detection and quantification (cognition)	Signal data (EEG)	Local	118 PD	SVM, kNN
García et al. [48]	3.3	Symptom detection and quantification (cognition)	Voice data	Local	40 PD, 40 HC	SVM, AdaBoost
Morales et al. [49]	3.3	Symptom detection and quantification (cognition)	Imaging data (MRI)	Local	16 PD-CN, 15 PD-MCI, 14 PDD	NB, SVM
Shibata et al. [50]	3.3	Symptom detection and quantification (cognition)	Imaging data (MRI)	Local	61 PD-MCI, 59 PD-CN	RF, gradient boosting, light gradient boosting
Hannink et al. [51]	3.4	Symptom detection and quantification (gait)	Motion data (gait)	eGaIT	99 geriatric subjects	CNN
Lu et al. [52]	3.4	Symptom detection and quantification (gait)	Motion data (gait)	Local	55 PD	CNN
Gao et al. [53]	3.4	Symptom detection and quantification (falls)	Clinical data (motor exams)	Local	148 PD (45 fallers, 97 non-fallers); 103 PD (41 fallers, 62 non-fallers)	LR, SVM, RF, gradient boosting
Severson et al. [54]	4.1	Disease progression (disease states)	Clinical data	PPMI, local	423 PD, 196 HC; 610 PD	HMM
Salmanpour et al. [55]	4.1	Disease progression (disease trajectories)	Clinical and imaging (MRI) data	PPMI	885 PD	k-means
Oxtoby et al. [56]	4.2	Disease progression (prediction of sequence of events)	Clinical and genetic data	Local, PPMI	100 PD, 33 HC; 350 PD, 127 HC	Event-based model
Latourelle et al. [57]	4.2	Disease progression (prediction of motor progression)	Clinical and genetic data	PPMI, local	312 PD, 117 HC; 317 PD	REFS

(continued)

Table 3
(continued)

Study	Subheading	Objectives/tasks	Modalities	Data sets	Number of subjects	Methods
Ahmadi Tastegar et al. [58]	4.2	Disease progression (prediction of motor and non-motor symptoms)	Clinical data and inflammatory cytokine measurements	LRRK2 cohort consortium	80 iPD, 80 PD-LRRK2	Elastic-net, RF
Amara et al. [59]	4.2	Disease progression (prediction of excessive daytime sleepiness)	Clinical and imaging (DaTscan) data	PPMI	423 PD, 196 HC	Random survival forest
Couronné et al. [60]	4.2	Disease progression (prediction of motor and non-motor symptoms)	Clinical and imaging (DaTscan) data	PPMI	362 PD	Generative mixed effect model
Faouzi et al. [61]	4.2	Disease progression (prediction of impulse control disorders)	Clinical and genetic data	PPMI, local	380 PD; 388 PD	LR, SVM, RF, gradient boosting, RNN
Yang et al. [63]	5.1	Treatment adjustment (motor symptoms)	Imaging data (fMRI)	Local	38 PD	Ridge, SVM, AdaBoost, gradient boosting
Kim et al. [64]	5.1	Treatment adjustment (motor symptoms)	Clinical and medication data	PPMI	431 PD	Markov decision process
Boutet et al. [66]	5.2	Optimal deep brain stimulation parameters	Imaging data (fMRI)	Local	67 PD	LDA
Geraedts et al. [67]	5.2	Optimal deep brain stimulation parameters	Signal data (EEG)	Local	112 PD	RF

Phokaewvarangkul et al. [68]	5.3	Electrical muscle stimulation (resting tremor)	Motion data (hand tremor)	Local	20 PD	LR, RF, SVM, MLP, LSTM
Panyakaew et al. [69]	5.3	Adverse event prevention (identification of modifiable risk factors of falls)	Clinical data	Local	305 PD	Gradient boosting

Modalities: CSF cerebrospinal fluid, DaTscan dopamine transporter scan, EEG electroencephalography, FDG-PET [¹⁸F]-fluorodeoxyglucose positron emission tomography, fMRI functional magnetic resonance imaging, LFP local field potential, MRI magnetic resonance imaging, PET positron emission tomography, rs-EEG resting-state electroencephalography, rs-fMRI resting-state functional magnetic resonance imaging

Data sets: eGait embedded Gait analysis using Intelligent Technologies (<https://www.mad.ti.fau.de/research/activitynet/digital-blobank/>), IPVC Italian Parkinson's Voice and Speech (<https://ieece-datasport.org/open-access/italian-parkinsons-voice-and-speech>), LRRK2 Cohort Consortium (<https://www.michaeljfox.org/news/lrrk2-cohort-consortium>), PPMI Parkinson's Progression Markers Initiative (<https://www.ppmi-info.org>)

Subjects: APS atypical parkinsonian syndromes, CP chronic pain, FOG freezing of gait, HC healthy controls, iPD idiopathic Parkinson's disease, iRBD idiopathic rapid eye movement sleep behavior disorder, MDD major depressive disorder, MSA-C cerebellar variant of multiple system atrophy, MSA-P parkinsonian variant of multiple system atrophy, PD Parkinson's disease, PD-D Parkinson's disease with normal cognition, PD-HYH Parkinson's disease with Hoehn and Yahr stage 2, PD-HYIII Parkinson's disease with Hoehn and Yahr stage 3, PD-LRRK2 Parkinson's disease with LRRK2 mutation, PD-MCI Parkinson's disease with mild cognitive impairment, PSP progressive supranuclear palsy

Methods: BNN Bayesian neural network, CNN convolutional neural network, DT decision tree, HMM hidden Markov model, JESS joint feature-sample selection, KFD kernel Fisher discriminant, kNN k-nearest neighbors, LR logistic regression, LDA linear discriminant analysis, LSTM long short-term memory, MFCC-GMM Mel-frequency cepstral coefficients-Gaussian mixture model, MLP multi-layer perceptron, NB naive Bayes, REFS reverse engineering and forward simulation, RF random forest, RNN recurrent neural network, RUSBoost random under-sampling boosting, SVM support vector machine

machines (SVM), and random forests (RF). They obtained the best predictive performance with the random forest trained on features normalized using multiple regression.

Mirelman and colleagues [11] also investigated gait and mobility measures that are indicative of PD and PD stages. They gathered data from sensors adhered to the participant's lower back, bilateral ankles, and wrists, during short walks, and extracted gait features. They investigated several strategies to perform feature selection and use a random under-sampling boosting classification algorithm to tackle class imbalance. When comparing PD patients with mild PD severity (Hoehn and Yahr stage 1) to healthy controls, they obtained good discriminative performance (84% sensitivity, 80% specificity). Most discriminative features were extracted from the upper limb sensors, with the remaining features extracted from the trunk sensor, while the lower limb sensors did not contribute to discrimination accuracy.

Kostikis and colleagues [12] investigated upper limb tremor using a smartphone-based tool. Signals from the phone's accelerometer and gyroscope were computed, from which features were extracted. They trained several machine learning algorithms, including random forest, naive Bayes, logistic regression (LR), and support vector machine, using these features as input and obtained the highest discriminative performance between PD patients and HC with the random forest model.

Kotsavasiloglou and colleagues [13] investigated the use of a pen-and-tablet device to study the differences in hand movement and muscle coordination between PD patients and HC. Data consisted of the trajectory of the pen's tip and on the pad's surface from drawings of simple horizontal lines, from which they extracted features. They investigated several machine learning algorithms, such as logistic regression, support vector machine, and random forest, and used nested cross-validation to perform feature selection. They obtained the highest discriminative performance with the naive Bayes model.

2.1.2 PD Diagnosis Using Voice Data

Voice data is usually recorded from high-quality microphones or from smartphones during specific vocal tasks focused on characteristics such as phonation and speech. Features are then extracted from the corresponding signals and used as input to machine learning classification algorithms.

Amato and colleagues [14] analyzed specific phonetic groups in native Italian speakers, extracted several spectral moments from the signals, and trained a SVM algorithm on these extracted features to distinguish PD patients from HC. They first worked on a public data set called Italian Parkinson's Voice and Speech,¹ with data

¹ <https://iee-e-dataport.org/open-access/italian-parkinsons-voice-and-speech>

recorded in ideal publications, and obtained great performance on the validation and test sets. They then merged this public data set with a data set that they collected, with data being recorded in more realistic, suboptimal conditions, and obtained good but lower performance on the validation and test sets of this merged data set. Experiments with training on one single data set and validation on the other data set were not performed, but it would have been interesting to estimate how well a trained model could generalize on other data sets with data being recorded in different conditions.

Jeancolas and colleagues [15] investigated the early diagnosis of PD and possible gender differences in voice data. They used a pre-trained deep neural network focused on speaker recognition system to extract features and obtained a higher performance than with a standard multidimensional Gaussian mixture model, although the increase was more important among men than women. They also investigated the impact of the quality of the recordings (using either a high-quality microphone or a telephone) and obtained the same conclusions in both cases.

In another study, Jeancolas and colleagues [16] investigated the differentiation between early PD patients and patients with idiopathic rapid eye movement sleep behavior disorders (iRBD), which are important risk factors to develop PD in the near future. They extracted features related to prosody, phonation, speech fluency, and rhythm abilities from speech recordings. They once again obtained a higher predictive performance among men than women in the PD vs HC classification task and a better discriminative power for this classification task than for the iRBD vs HC one, suggesting that discriminating iRBD patients from HC using voice data is a much harder task, but it is also probably a most useful one in practice.

Quan and colleagues [17] investigated the extraction of global static features (from the whole signals) and local dynamic features (using a sliding window on the signals) from voice data during articulation tasks. They trained standard machine learning classification algorithms, such as decision trees (DT), k-nearest neighbors, naive Bayes, and support vector machines, using the static features, while they trained a recurrent neural network, more specifically a bidirectional long short-term memory (LSTM), on the dynamic features and obtained a higher predictive performance with the deep learning approach.

Although many studies reported high predictive performances, some results must be taken with caution. Indeed, a recent study reported methodological issues in several studies, including record-wise cross-validation instead of subject-wise cross-validation, high imbalance in ages between PD patients and HC, and performance metrics computed on the validation folds of k-fold cross-validation and not on an independent test set, which may lead to overly optimistic results [18].

2.1.3 PD Diagnosis Using Imaging Data

The diagnosis of PD remains based on its clinical presentation [19]. Imaging of dopaminergic terminals loss can be assessed using nuclear imaging, but it is not recommended in clinical routine and does not differentiate PD from other related disorders associated with dopamine neuron loss [20]. Standard brain magnetic resonance imaging (MRI) is normal in PD. However, several new markers have been recently been investigated in several studies, with mixed results.

Adeli and colleagues [21] investigated the use of T1-weighted anatomical MRI data to differentiate PD patients from HC. They developed a joint feature-sample selection algorithm in order to select an optimal subset of both features and samples from a training set, and a robust classification framework that performs denoising of the selected features and samples then learns a classification model. They analyzed data from 374 PD patients and 169 HC from the Parkinson's Progression Markers Initiative² (PPMI) cohort and included white matter, gray matter, and cerebrospinal fluid measurements from 98 regions of interest. The combination of the proposed feature selection/extraction method and classifier achieved the highest predictive accuracy (0.819), being significantly better than almost every other combination of a feature selection/extraction method and a classification algorithm.

Solana-Lavalle and Rosas-Romero [22] investigated the use of voxel-based morphometry features extracted from T1-weighted anatomical MRI to perform a PD vs HC classification task. Their pipeline consisted of five stages: (i) identification of regions of interest using voxel-based morphometry, (ii) analysis of these regions for PD detection, (iii) feature extraction based on first- and second-order statistics, (iv) feature selection based on principal component analysis, and (v) classification with tenfold cross-validation based on seven different algorithms (including k-nearest neighbors, support vector machine, random forest, naive Bayes, and logistic regression). They obtained excellent predictive performance for both male and female genders and for both 1.5 T and 3 T MRI scans (accuracy scores ranging from 0.93 to 0.99 for the best classification algorithms). However, cross-validation was performed very late in their pipeline (after the feature subset selection), which could lead to biased models and overly optimistic predictive performances.

Mudali and colleagues [23] investigated another modality, [¹⁸F]-fluorodeoxyglucose positron emission tomography (FDG-PET), to compare 20 PD patients and 18 HC. They applied the subprofile model/principal component analysis method to extract features from the images. They considered a DT algorithm and used leave-one-out cross-validation to evaluate the predictive

² <https://www.ppmi-info.org>

performance of the models. They obtained really low predictive performance (50% sensitivity, 45% specificity), close to chance level.

Overall, it is unclear if machine learning applied to anatomical MRI or FDG-PET can bring added value for the diagnosis of PD. However, advanced MRI sequences have the potential to bring much more valuable information [24].

2.2 Differential Diagnosis

The PD vs HC binary classification task has limited utility as, even at the early stage of PD, patients have clinical symptoms strongly suggesting that they suffer from a movement disorder and thus are not healthy subjects. However, the accurate early diagnosis of parkinsonian syndromes is difficult but needed due to the different pathologies and thus the different care. Although one study investigated the differential diagnosis using sensor-based gait analysis [25], most studies investigated it using imaging data, particularly diffusion MRI.

Huppertz and colleagues [26] investigated the differential diagnosis with data from a relatively large cohort (73 HC, 204 PD, 106 PSP, 20 MSA-C, and 60 MSA-P). Using atlas-based volumetry of brain MRI data, they extracted volumes in several regions of interest and trained and evaluated a linear SVM algorithm using leave-one-out cross-validation. They obtained good predictive performance in most binary classification tasks and showed that midbrain, basal ganglia, and cerebellar peduncles were the most relevant regions.

A landmark study on this topic was published in 2019 by Archer and colleagues [27], with diffusion-weighted MRI data being collected for 1002 subjects from 17 MRI centers in Austria, Germany, and the USA. They extracted 60 free-water and 60 free-water-corrected fractional anisotropy values from diffusion-weighted MRI data, and the other features consisted of the third part of the Movement Disorder Society-Sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS III), sex, and age. They trained several SVM models and showed that the model trained using MDS-UPDRS III (with sex and age also, for all the models) performed poorly in most classification tasks, whereas the model trained using DWI features had much higher predictive performance (particularly for the MSA vs PSP task), and adding MDS-UPDRS III to this model did not improve the performance.

More recently, Chougar and colleagues [28] investigated the replication of such differential diagnosis models in clinical practice on different MRI systems. Using MRI data from 119 PD, 51 PSP, 35 MSA-P, 23 MSA-C, and 94 HC, split into a training cohort ($n = 179$) and a replication cohort ($n = 143$), they extracted volumes and diffusion tensor imaging (DTI) features (fractional anisotropy, mean diffusivity, axial diffusivity, and radial diffusivity) in 13 regions of interest. They investigated two feature normalization strategies (one based on the data of all subjects in the training

set and one based on the data of HC for each MRI system to tackle the different feature distributions, in particular for DTI features, because of the use of different MRI systems) and four standard machine learning algorithms, including logistic regression, support vector machines, and random forest. They obtained high performances in the replication cohort for many binary classification tasks (PD vs PSP, PD vs MSA-C, PSP vs MSA-C, PD vs atypical parkinsonism), but lower performances for other classification tasks involving MSA-P patients (PD vs MSA-P, MSA-C vs MSA-P). They showed that adding DTI features did not improve performance compared to using volumes only and that the usual normalization strategy worked best in this case.

Shinde and colleagues [29] investigated the automatic extraction of contrast ratios of the substantia nigra pars compacta from neuromelanin-sensitive MRI using a convolutional neural network. Based on the class activation maps, they identified that the left side of substantia nigra pars compacta played a more important role in the decision of the model compared to the right side, in agreement with the concept of asymmetry in PD.

A recent study [30] investigated the use of positron emission tomography of the translocator protein, expressed by glial cells, and extracted normalized standardized uptake value images and normalized total distribution volume images. Using a linear discriminant analysis algorithm with leave-one-subject-out cross-validation, they obtained great discriminative power between MSA and PD patients, with better performance with normalized total distribution volume images.

2.3 Disease Understanding

Rather than focusing on the diagnosis of Parkinson's disease itself, several studies were more focused on interpreting the trained machine learning models in order to better understand the mechanisms of Parkinson's disease.

Khawaldeh and colleagues [31] investigated the task-related modulation of local field potentials of the subthalamic nucleus before and during voluntary upper and lower limb movements in 18 consecutive Parkinson's disease patients undergoing deep brain stimulation (DBS) surgery of the subthalamic nucleus in order to improve motor symptoms. Using a naive Bayes classification algorithm, they obtained chance-level performance at rest, but much higher performance during the pre-cue, pre-movement onset, and post-movement onset tasks. They showed that the presence of bursts of local field potential activity in the alpha and, even more so, in the beta frequency band significantly compromised the prediction of the limb to be moved, concluding that low-frequency bursts restrict the capacity of the basal ganglia system to encode physiologically relevant information about intended actions.

Poston and colleagues [32] investigated brain mechanisms that allow some PD patients with severe dopamine neuron loss to remain cognitively normal. Using functional MRI data from PD patients without cognitive impairment and from HC collected during a working memory task, they trained a support vector machine classifier and identified robust differences in putamen activation patterns, providing novel evidence that PD patients maintain normal cognitive performance through compensatory hyperactivation of the putamen.

Trezzini and colleagues [33] investigated cerebrospinal fluid biomarkers, and more precisely the metabolome, in early-stage PD. The logistic regression model trained on such data provided good discriminative power, and the most associated biomarkers were mannose, threonic acid, and fructose. These biomarkers were associated with antioxidative stress response, glycation, and inflammation and may help better understand PD pathogenesis.

Vanneste and colleagues [34] investigated thalamocortical dysrhythmia, which is a model proposed to explain divergent neurological disorders and is characterized by a common oscillatory pattern in which resting-state alpha activity is replaced by cross-frequency coupling of low- and high-frequency oscillations. The trained support vector machine model identified specific brain regions that provided good discriminative power between PD patients and HC, including subgenual anterior cingulate cortex, posterior cingulate cortex, parahippocampus, dorsal anterior cingulate cortex, and motor cortex. Another model also identified brain areas that are common to the pathology of Parkinson's disease, pain, tinnitus, and depression, including dorsal anterior cingulate cortex and parahippocampal area.

3 Symptom Detection and Quantification

Given the complexity and heterogeneity of Parkinson's disease, prompt accurate assessment of symptoms is needed. A detailed scale, called the Movement Disorder Society-Sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [35], is currently the gold standard to assess motor (and non-motor) features of PD patients by movement disorder specialists. The scale is divided into four sections. The first two sections allow for assessing the non-motor and motor activities of daily living, respectively, while the third section consists of a motor exam, and the fourth section allows for assessing motor complications.

Nonetheless, the MDS-UPDRS has several limitations. First, it requires time (30–45 minutes for the full scale) and a trained movement disorder specialist to fill it, limiting its use during clinical routine visits. Second, part of subjectivity from a human evaluation,

and thus variance in the MDS-UPDRS scores, cannot be excluded, with a recent study suggesting that MDS-UPDRS scores contain a substantial amount of variance [36]. Moreover, other scales are typically used to more precisely assess non-motor symptoms such as depression, anxiety, and cognition. Finally, scales are addressed during a visit at the hospital and may not reflect the symptoms in a more ecological setting, at home, during the daily life of the patient. Automatic detection and quantification of symptoms using machine learning may help tackle these limitations, and several studies investigated this topic. In the remaining of this section, we group these studies based on the symptoms investigated.

3.1 Freezing of Gait

Freezing of gait (FOG) is a common motor symptom and is associated with life-threatening accidents such as falls. Prompt identification or prediction of freezing of gait episodes is thus needed.

Ahlrichs and colleagues [37] investigated freezing of gait in 20 PD patients (8 with FOG, 12 without FOG), split into a training set (15 patients) and a test set (5 patients). They collected sensor (accelerometer, gyroscope, and magnetometer) data during scripted activities (e.g., walking around the apartment, carrying a full glass of water from the kitchen to another room) and non-scripted activities (e.g., answering the phone). Two recording sessions were considered, one in “OFF” motor state and one in “ON” motor state, and the data was labeled by experienced clinicians based on the corresponding video recordings. The task was a binary classification task (FOG vs no FOG) for each window. They extracted sub-signals from the whole signals using a sliding window and then extracted features and in the time and frequency domains for each sub-signal. They trained two SVM algorithms (one with a linear kernel, one with a Gaussian kernel) and obtained high and better results with the linear kernel.

Aich and colleagues [38] gathered sensor data for 36 PD patients with FOG and 15 PD patients without FOG from 2 wearable triaxial accelerometers during clinical experiments. They extracted features, such step time, stride time, step length, stride length, and walking speed, from the signals. They trained several classic machine learning classification algorithms (SVM, kNN, DT, NB) and obtained good predictive performances with all of them, although the SVM model had the highest mean accuracy on the test sets of the cross-validation procedure.

Borzi and colleagues [39] collected data from 2 inertial sensors placed on each shin of the 11 PD patients during the “timed up and go” test in order to investigate FOG and pre-FOG detection. They extracted features in the time and frequency domains and trained decision tree algorithms. They obtained great predictive performance to detect FOG episodes, but lower performance to predict pre-FOG episodes, with the performance decreasing even more as the window length increased.

Dvorani and colleagues [40] were interested in detecting foot motion phases using a shoe-placed inertial sensor in order to detect FOG episodes. They extracted ten features, including stride length, maximum gait velocity, and step duration, from each motion phase and trained a SVM algorithm to detect FOG episodes. They obtained great performance when using features from the current and two preceding motion phases, but lower performance when using only features from the two preceding motion phases, highlighting the higher difficulty to predict FOG episodes in advance. Shalin and colleagues [41] reached the same conclusion using plantar pressure data and a long short-term memory neural network.

3.2 Bradykinesia and Tremor

Bradykinesia and tremor are two other motor symptoms that are frequently investigated for automatic assessment.

Park and colleagues [42] investigated automated rating for resting tremor and bradykinesia from video clips of resting tremor and finger tapping of the bilateral upper limbs. They extracted several features from the video clips, including resting tremor amplitude and finger tapping speed, amplitude, and fatigue, using a pre-trained deep learning model. These features were used as input of a SVM algorithm to predict the corresponding scores from the MDS-UPDRS scale. For resting tremors, the automated approach had excellent reliability range with the gold standard rating and higher performance than that of non-trained human rater. For finger tapping, the automated approach had good reliability range with the gold standard rating and similar performance than that of non-trained human rater.

Kim and colleagues [43] performed a study in which they investigated tremor severity using three-dimensional acceleration and gyroscope data obtained from wearable device. They investigated a convolutional neural network to automatically extract features and perform classification, compared to extracting defined features from the time and frequency domains and training standard machine learning algorithms (random forest, naive Bayes, linear regression, support vector machines) using these features. They obtained better higher predictive performance with the deep learning approach than the standard machine learning approach. Eskofier and colleagues [44] obtained similar results using inertial measurement units collected during motor tasks.

3.3 Cognition

Cognitive impairment is frequent in PD, with the point prevalence of PD dementia being around 30% and the cumulative prevalence for patients surviving more than 10 years being at least 75% [45]. Due to its high negative impact on the quality of life of PD patients and their caregivers, it is important to identify and quantify cognitive impairment. Several scales to assess cognition already

exist, such as the Mini-Mental State Examination and the Montreal Cognitive Assessment, but automatic assessment of cognition could be helpful.

Abós and colleagues [46] investigated discriminating cognitive status in PD through functional connectomics. Using resting-state functional MRI data, they extracted features consisting of connection-wise pattern of functional connectivity. They performed feature selection using randomized logistic regression with leave-one-out cross-validation and then trained a SVM algorithm. They obtained good discriminative performance between PD patients with mild cognitive impairment and with no cognitive impairment, but could not report significant connectivity reductions between both groups.

Betrouni and colleagues [47] investigated the use of electroencephalograms to automatically assess their cognitive status. A cluster analysis of the neuropsychological assessments of 118 PD patients revealed 5 cognition clusters. They extracted quantitative features from the electroencephalograms and performed feature selection based on Pearson correlation tests. They trained two machine learning algorithms (kNN and SVM), using a fivefold cross-validation procedure that was repeated five times, and obtained good similar predictive performances for the five-class classification task with both models.

García and colleagues [48] investigated cognitive decline using dysarthric symptoms. They extracted prosodic, articulatory, and phonemic identifiability features from speech signals recorded during the reading of two narratives. Using a SVM algorithm and nested cross-validation, they obtained correct discriminative performance (area under the receiver operating characteristics curve of 0.76), with the highest performance being obtained using phonemic identifiability features.

Morales and colleagues [49] investigated the classification of PD patients with no cognitive impairment ($n = 16$), with mild cognitive impairment ($n = 15$), and with dementia ($n = 14$). They trained several variants of the naive Bayes algorithm and 1 SVM algorithm on 112 MRI features consisting of volumes of subcortical structures and thickness of cortical parcels and obtained good discriminative performance in the 3 binary classification tasks, the lower performance corresponding to the differentiation between PD patients with no cognitive impairment with mild cognitive impairment. The most important features involved the following brain regions: left cerebral cortex, left caudate, left entorhinal, right inferior left hippocampus, and brainstem.

A recent study [50] also investigated MRI data, more specifically quantitative susceptibility mapping images parcellated into 20 regions of interest, for the early detection of cognitive impairment in PD. Using tree-based ensemble machine learning algorithms, such as random forest and extreme gradient boosting,

they obtained acceptable predictive performance and showed that the features corresponding to the caudate nucleus were important for classification and also inversely correlated with Montreal Cognitive Assessment scores.

3.4 Other Symptoms

Although less prevalent in the literature, studies also investigated other PD symptoms such as falls and motor severity.

An early study by Hannink and colleagues [51] was performed to investigate gait parameter extraction from sensor data using convolutional neural networks. Using 3d-accelerometer and 3d-gyroscope data from 99 geriatric patients, the objective was to predict the stride length and width, the foot angle, and the heel and toe contact times. They investigated two approaches to tackle this multi-output regression task, either training a single convolutional neural network to predict the five outcomes or training a convolutional neural network for each outcome, and obtained better performance on an independent test set with the latter approach. Although the considered population was not parkinsonian, the prevalence of gait symptoms in this population and the obtained results might be relevant to better understand gait in this population. Lu and colleagues [52] investigated gait in PD, as measured by MDS-UPDRS item 3.10, which does not include freezing of gait. They collected video recordings of MDS-UPDRS exams from 55 participants which were scored by 3 different trained movement disorder neurologists, and the ground truth score was defined using majority voting among the 3 raters. They performed skeleton extraction from the videos and trained a convolutional neural network, with regularization using rater confusion estimation to tackle noise in labels, to predict gait severity. They obtained correct performance on the test set (72% accuracy with majority voting, 84% accuracy with the model predicting at least one of the raters' scores).

Gao and colleagues [53] investigated falls in two data sets independently collected at two different sites. Using clinical scores as input, they trained several classic machine learning classification algorithms to differentiate fallers from non-fallers. They obtained acceptable predictive performance in both data sets when training and evaluating (using cross-validation) a model in each data set independently. They also showed that the predictive performance was lower when training the model on one data set and evaluating it on the other data set, which is not surprising, but it is important to have this possible issue in mind when a model is not evaluated on a different cohort.

4 Disease Progression

Given the complexity and heterogeneity of Parkinson's, prediction of disease progression with individual trajectories is challenging. Two subtypes of PD, one with more postural instability and gait difficulty and the other one with more tremor symptoms, are already known. Nonetheless, there are other motor symptoms in PD, and many PD symptoms are non-motor; thus, deeper knowledge is required to understand disease progression.

4.1 Disease Subtypes

Several studies focused on the identification of more specific disease subtypes than the two aforementioned well-known ones characterized by postural instability and gait difficulty for one and tremor predominant for the other.

Severson and colleagues [54] worked on the development of a statistical progression model of Parkinson's disease accounting for intra-individual and inter-individual variability, as well as medication effects. They built a contrastive latent variable model followed by a personalized input-output hidden Markov model to define disease states and assessed the clinical significance of the states on seven key motor or cognitive outcomes (mild cognitive impairment, dementia, dyskinesia, presence of motor fluctuations, functional impairment from motor fluctuations, Hoehn and Yahr score, and death). They identified eight disease states that were primarily differentiated by functional impairment, tremor, bradykinesia, and neuropsychiatric measures. The terminal state had the highest prevalence of key clinical outcomes, including almost every recorded instance of dementia. The discovered states were non-sequential, with overlapping disease progression trajectories, supporting the use of non-deterministic disease progression models, and suggesting that static subtype assignment might be ineffective at capturing the full spectrum of PD progression.

Salmanpour and colleagues [55] performed a longitudinal clustering analysis and prediction PD progression. They extracted almost a thousand features, including motor, non-motor, and radiomics features. They performed a cross-sectional clustering analysis and identified three distinct progression trajectories, with two trajectories being characterized by disease escalation and the other trajectory by disease stability. They also investigated the prediction of progression trajectories from early stage (baseline and year 1) data and obtained the highest predictive performance with a probabilistic neural network.

4.2 Prediction of Future Motor and Non-motor Symptoms

Prediction of future symptoms and individual disease trajectories was the main focus of several studies.

Oxtoby and colleagues [56] aimed at estimating the sequence of clinical and neurodegeneration events, and variability in this

sequence, using data-driven disease progression modelling, with a focus on PD patients with higher risk of developing dementia (defined as PD patients being diagnosed at age 65 or later). They analyzed baseline visit data from two separate cohorts: a local discovery cohort (100 PD patients and 33 HC) and a replication cohort (PPMI study, 350 PD patients and 127 HC). They considered 42 features, including 8 clinical/cognitive measures, 6 vision measures, 4 retinal measures, 8 regional measures of cortical thickness, 4 measures of white matter neurodegeneration in the substantia nigra, and 12 regional measures of brain iron content. They trained event-based models that incorporate non-parametric mixture modelling using ten fivefold cross-validation procedures to estimate the robustness of the models. The authors showed that Parkinson's progression in patients at higher risk of developing dementia starts with classic prodromal features of PD (sleep and olfactory disorders), followed by early deficits in visual abilities and increased brain iron content, followed later by a less certain ordering of neurodegeneration in the substantia nigra and cortex, neuropsychological cognitive deficits, retinal thinning in dopamine layers, and further visual deficits. Their results support the growing piece of evidence that visual processing specifically is affected early in PD patients with high risk of developing dementia.

Latourelle and colleagues [57] investigated the development of predictive models of motor progression using longitudinal clinical, molecular, and genetic data. More specifically, the objective was to predict the annual rate of changes in combined scores from the second and third parts of the MDS-UPDRS. The trained model showed strong performance in the training cohort (using fivefold cross-validation) and lower but still significant performance in an independent replication cohort. The most relevant features included baseline MDS-UPDRS motor score, sex, and age, as well as a novel PD-specific epistatic interaction. Genetic variation was the most useful prediction of motor progression, and baseline CSF biomarkers had a lower but still significant effect on predicting motor progression. They also performed simulations with the trained model and concluded that incorporating the predicted rates of motor progression into the final models of treatment effect reduced the variability in the study outcome, allowing significant differences to be detected at sample sizes up to 20% smaller than in naive trials.

Ahmadi Rastegar and colleagues [58] investigated the prediction of longitudinal clinical outcomes after 2-year follow-up from baseline and 1-year follow-up data. They also measured 27 inflammatory cytokines and chemokines in serum at baseline and after 1 year to investigate cytokine stability. Training random forest algorithms, the best prediction models were for motor symptom severity scales (Hoehn and Yahr stage and MDS-UPDRS III total score), and several inflammatory cytokine and chemokine features

were among the most relevant features to predict Hoehn and Yahr stage and MDS-UPDRS III total score, giving evidence that peripheral cytokines may have utility for aiding prediction of PD progression using machine learning models.

Amara and colleagues [59] investigated the prediction of future incidents of excessive daytime sleepiness. They trained a random survival forest using 33 baseline variables, including anxiety, depression, rapid eye movement sleep, cognitive scores, α -synuclein, p-tau, t-tau, and ApoE e4 status. The performance of the model was only marginally better than random guess, but the strongest predictive features were p-tau and t-tau.

Couronné and colleagues [60] performed longitudinal data analysis to predict patient-specific trajectories. They proposed to use a generative mixed effect model that considers the progression trajectories as curves on a Riemannian manifold and that can handle missing values. They applied their model to PD progression with joint modelling of two features (MDS-UPDRS III total score and striatal binding ration in right caudate). Interpretation of the model revealed that patients with later onset progress significantly faster and that α -synuclein mean level was correlated with PD onset.

Faouzi and colleagues [61] investigated the prediction of future impulse control disorders (psychiatric disorders characterized by the inability to resist an urge or an impulse and which include a wide range of types including compulsive shopping, internet addiction, and hypersexuality, for instance) in Parkinson's disease. The objective of their study was to predict the presence or absence of these disorders at the next clinical visit of a given patient. Using clinical and genetic data, they trained several machine learning models on a training cohort and evaluated the models on the training cohort (using cross-validation) and on an independent replication cohort. They showed that a recurrent neural network model achieved significantly better performance than a trivial model (predicting the status at the next visit with the status at the most recent visit), but the increase in performance was too small to be deemed clinically relevant. Nevertheless, this proof-of-concept study highlights the potential of machine learning for such prediction.

5 Treatment Adjustment and Adverse Event Prevention

Being able to predict future adverse events in Parkinson's disease is useful, but being able to prevent them would be even more useful. Parkinson's disease is one of the few neurodegenerative diseases where current therapies can greatly improve the quality of life of the patients, but these therapies also have adverse effects. Providing personalized adapted therapies to every patient is of high importance.

Machine learning allows for unveiling complex correlations or patterns from data. However, correlation does not imply causality: if two variables are correlated, one variable does not necessarily cause an effect on the other. Therefore, standard machine learning is not always well adapted to draw conclusions for personalized therapies. Ultimately, clinical trials with a specific hypothesis tested are the best solution to draw causality effect conclusions. Nonetheless, several machine learning approaches can investigate causality effects. Causal inference, that is, being able to discover which variables have which impacts on which other variables, is an open research topic in machine learning, but usually requires a lot of data, limiting its use in Parkinson's disease. Nonetheless, exploratory studies suggesting potential options for personalized therapies and adverse event prevention have been published.

5.1 Dopamine Replacement Therapy

Dopamine replacement therapy, as a way to compensate the loss of dopamine neurons in the brain, is the most common therapy due to its efficacy and simplicity (drug intake). Nonetheless, it also comes with adverse effects and long-term motor complications such as motor fluctuations (worsening or reappearance of motor symptoms before the next drug intake) and dyskinesia (involuntary muscle movements) [62].

Yang and colleagues [63] investigated the utility of amplitude of low-frequency fluctuation computed from functional MRI data of 38 PD patients in order to predict individual patient's response to levodopa treatment. They applied principal component analysis to perform dimensionality reduction and trained gradient tree boosting algorithms to discriminate between moderate and superior responders to levodopa treatment. Treatment efficacy was defined based on motor symptom improvement from the state of medication off to medication on, as assessed by MDS-UPDRS III total score. They obtained great discriminative performance between both groups, even though no significant difference in clinical data was observed between both groups. The mainly contributed regions for both models included the bilateral primary motor cortex, the occipital cortex, the cerebellum, and the basal ganglia. These results suggest the potential utility of amplitude of low-frequency fluctuation as promising predictive markers of dopaminergic therapy response in PD patients.

Kim and colleagues [64] investigated the use of reinforcement learning to predict optimal treatment for reducing motor symptoms. They derived clinically relevant disease states and an optimal combination of medications for each of them by using policy iteration of the Markov decision process. Their model achieved a lower level of motor symptom severity scores than what clinicians did, whereas the clinicians' medication rules were more consistent than their model. Their model followed the clinician's medication rules in most cases but also suggested some changes, which leads to the

difference in lowering symptom severity. This proof of concept showed the potential utility of reinforcement learning to derive optimal treatment strategies.

5.2 Deep Brain Stimulation

Deep brain stimulation is a neurosurgical procedure that uses implanted electrodes and electrical stimulation and has proven efficacy in advanced Parkinson's disease by decreasing motor fluctuations and dyskinesia and improving quality of life [65]. The most commonly stimulated region is the subthalamic nucleus, but the globus pallidus is sometimes preferred. Although DBS usually greatly improves the motor symptoms, it also has downsides, such as requiring personalized parameters and potential adverse events such as postoperative cognitive decline.

Boutet and colleagues [66] investigated the prediction of optimal deep brain stimulation parameters from functional MRI data. They extracted blood-oxygen-level-dependent (BOLD) signals in 16 motor and non-motor regions of interest for 67 PD patients, from which 62 underwent DBS of the subthalamic nucleus and 5 underwent DBS of the globus pallidus. They trained a linear discriminant analysis algorithm on normalized BOLD changes using fivefold cross-validation and obtained great performance in classifying optimal vs non-optimal parameter settings, although the performance was lower on two additional (a priori clinically optimized and in stimulation-naïve patients) unseen data sets.

Geraedts and colleagues [67] also investigated deep brain stimulation in the context of cognitive function, as a downside of DBS for PD is the potential deterioration of cognition postoperatively. They extracted features from electroencephalograms, trained random forest algorithms using tenfold cross-validation, and obtained great discrimination between PD patients with the best and worst cognitive performances. However, it should be noted that they only included the best and worst cognitive performers ($n = 20$ per group from 112 PD patients), making the classification task probably much easier than if it was performed on the 112 PD patients, thus requiring their model to be evaluated on PD patients independently on their cognitive performance. Nonetheless, their results suggest the potential utility of electroencephalography for cognitive profiling in DBS.

5.3 Others

Phokaewvarangkul and colleagues [68] explored the effect of electrical muscle stimulation as an adjunctive treatment for resting tremor during "ON" period, with machine learning used to predict the optimal stimulation level that will yield the longest period of tremor reduction or tremor reset time. They used sensor data from a glove incorporating a three-axis gyroscope to measure tremor signals. The stimulation levels were discretized into five ordinary classes, with the objective to predict the accurate class from the sensor data. They observed a significant reduction in tremor

parameters during stimulation. The best performing machine learning model was a LSTM neural network in comparison to classic algorithms such as logistic regression, support vector machine, and random forest. The high predictive performance of the LSTM model confirmed the potential utility of electrical muscle stimulation for the reduction of resting tremors in PD.

Panyakaew and colleagues [69] investigated the identification of modifiable risk factors of falls. The input data consisted of clinical demographics, medications, and balanced confidence scaled by the 16-item Activities-Specific Balance Confidence (ABC-16) scale, from 305 PD patients (99 fallers, 58 recurrent fallers, and 148 non-fallers). They trained two gradient tree boosting algorithms using sevenfold cross-validation. They obtained good predictive performance at differentiating fallers from non-fallers, the most relevant features being item 7 (sweeping the floor), item 5 (reaching on tiptoes), and item 12 (walking in a crowded mall) from the ABC-16 scale, followed by disease stage and duration. They obtained even better performance at differentiating recurrent fallers from non-fallers, the most relevant features being items 12, 5, and 10 (walking across a parking lot) from the ABC-16 scale, followed by disease stage and current age.

6 Conclusion

Many research works on Parkinson's disease and related disorders using machine learning have been published in the literature, investigating diagnosis, symptom severity, disease progression, and personalized therapies. These studies provide new insights to better understand these neurodegenerative disorders.

However, many questions and challenges are still open. The early-stage, and even more so the prodromal-stage, classification of Parkinson's disease is still very challenging. The early differential diagnosis of parkinsonian syndromes is another topic for which higher performance is needed at an early stage. More highly personalized therapies are also needed to better improve the quality of life of the PD patients. All the research works on these topics also need to be evaluated in non-research environments in order to be translated to the clinics.

Right usage of machine learning is required to try to answer these questions and challenges. The most common methodological issues are usually related to the cross-validation procedure used, which can lead to biased, overly optimistic, reported predictive performance. Nonetheless, our anecdotal experience after performing this literature review is that these issues are less and less frequent over time. Nonetheless, many studies use small data sets and leave-one-out cross-validation, which provides an unbiased estimation of the predictive performance, but with high variance.

The few studies that investigated replicating their results in another independent data set all reported (much) lower predictive performance, highlighting the critical need of replication. Using artificial intelligence algorithms also rises ethical and legal issues regarding the collection, processing, storage, and reuse of potentially sensitive patient data, particularly coming from sensor-based digital data [9]. These aspects will have to be taken into consideration when transferring results obtained from clinical research to clinical routine use.

To conclude, the use of machine learning has allowed researchers to better understand Parkinson's disease and related disorders and suggested potential to better diagnose these disorders as well as to provide better care for the patients, but more research works and replication studies are required to translate these results into the clinics through clinical decision support systems.

Acknowledgments

The authors would like to thank Jochen Klucken for his fruitful remarks. This work was supported by the French government under the management of Agence Nationale de la Recherche as part of the “Investissements d'avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6), by the European Union H2020 Programme (grant number 826421, project TVB-Cloud), and by the ERA PerMed EU-wide project DIGIPD (01KU2110).

References

1. GBD (2016) Parkinson's disease collaborators (2018) global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016. *Lancet Neurol* 17:939–953
2. Ascherio A, Schwarzschild MA (2016) The epidemiology of Parkinson's disease: risk factors and prevention. *Lancet Neurol* 15:1257–1272
3. Blauwendraat C, Nalls MA, Singleton AB (2020) The genetic architecture of Parkinson's disease. *Lancet Neurol* 19:170–178
4. Corti O, Lesage S, Brice A (2011) What genetics tells us about the causes and mechanisms of Parkinson's disease. *Physiol Rev* 91:1161–1218
5. Sambin S, Lavisce S, Decaix C et al (2022) Compensatory mechanisms nine years before Parkinson's disease conversion in a LRRK2 R1441H family. *Mov Disord* 37:428–430
6. Wenning GK, Stankovic I, Vignatelli L et al (2022) The Movement Disorder Society criteria for the diagnosis of multiple system atrophy. *Mov Disord* 37:1131
7. Tolosa E, Garrido A, Scholz SW et al (2021) Challenges in the diagnosis of Parkinson's disease. *Lancet Neurol* 20:385–397
8. Boxer AL, Yu J-T, Golbe LI et al (2017) Advances in progressive supranuclear palsy: new diagnostic criteria, biomarkers, and therapeutic approaches. *Lancet Neurol* 16:552–563
9. Fröhlich H, Bontridder N, Petrovska-Delacréta D et al (2022) Leveraging the potential of digital technology for better individualized treatment of Parkinson's disease. *Front Neurol* 13:788427
10. Wahid F, Begg RK, Hass CJ et al (2015) Classification of Parkinson's disease gait using spatial-temporal gait features. *IEEE J Biomed Health Inform* 19:1794–1802

11. Mirelman A, Ben Or Frank M, Melamed M et al (2021) Detecting sensitive mobility features for Parkinson's disease stages via machine learning. *Mov Disord* 36:2144–2155
12. Kostikis N, Hristu-Varsakelis D, Arnaoutoglou M et al (2015) A smartphone-based tool for assessing parkinsonian hand tremor. *IEEE J Biomed Health Inform* 19:1835–1842
13. Kotsavasiloglou C, Kostikis N, Hristu-Varsakelis D et al (2017) Machine learning-based classification of simple drawing movements in Parkinson's disease. *Biomed Signal Process Control* 31:174–180
14. Amato F, Borzì L, Olmo G et al (2021) Speech impairment in Parkinson's disease: acoustic analysis of unvoiced consonants in Italian native speakers. *IEEE Access* 9:166370–166381
15. Jeancolas L, Petrovska-Delacrétaz D, Mangone G et al (2021) X-vectors: new quantitative biomarkers for early Parkinson's disease detection from speech. *Front Neuroinform* 15:578369
16. Jeancolas L, Mangone G, Petrovska-Delacrétaz D et al (2022) Voice characteristics from isolated rapid eye movement sleep behavior disorder to early Parkinson's disease. *Parkinsonism Relat Disord* 95:86–91
17. Quan C, Ren K, Luo Z (2021) A deep learning based method for Parkinson's disease detection using dynamic features of speech. *IEEE Access* 9:10239–10252
18. Ozbolt AS, Moro-Velazquez L, Lina I et al (2022) Things to consider when automatically detecting Parkinson's disease using the phonation of sustained vowels: analysis of methodological issues. *Appl Sci* 12:991
19. Berg D, Adler CH, Bloem BR et al (2018) Movement disorder society criteria for clinically established early Parkinson's disease. *Mov Disord* 33:1643–1646
20. de la Fuente-Fernández R (2012) Role of DaTSCAN and clinical diagnosis in Parkinson disease. *Neurology* 78:696–701
21. Adeli E, Shi F, An L et al (2016) Joint feature-sample selection and robust diagnosis of Parkinson's disease from MRI data. *NeuroImage* 141:206–219
22. Solana-Lalalle G, Rosas-Romero R (2021) Classification of PPMI MRI scans with voxel-based morphometry and machine learning to assist in the diagnosis of Parkinson's disease. *Comput Methods Prog Biomed* 198:105793
23. Mudali D, Teune LK, Renken RJ et al (2015) Classification of parkinsonian syndromes from FDG-PET brain data using decision trees with SSM/PCA features. *Comput Math Methods Med* 2015:136921
24. Mitchell T, Lehéricy S, Chiu SY et al (2021) Emerging neuroimaging biomarkers across disease stage in Parkinson disease: a review. *JAMA Neurol* 78:1262–1272
25. Gaßner H, Raccagni C, Eskofier BM et al (2019) The diagnostic scope of sensor-based gait analysis in atypical parkinsonism: further observations. *Front Neurol* 10:5
26. Huppertz H-J, Möller L, Südmeyer M et al (2016) Differentiation of neurodegenerative parkinsonian syndromes by volumetric magnetic resonance imaging analysis and support vector machine classification. *Mov Disord* 31:1506–1517
27. Archer DB, Bricker JT, Chu WT et al (2019) Development and validation of the automated imaging differentiation in parkinsonism (AID-P): a multi-site machine learning study. *Lancet Digit Health* 1:e222–e231
28. Chougar L, Faouzi J, Pyatigorskaya N et al (2021) Automated categorization of parkinsonian syndromes using magnetic resonance imaging in a clinical setting. *Mov Disord* 36:460–470
29. Shinde S, Prasad S, Saboo Y et al (2019) Predictive markers for Parkinson's disease using deep neural nets on neuromelanin sensitive MRI. *Neuroimage Clin* 22:101748
30. Jucaite A, Cselényi Z, Kreisl WC et al (2022) Glia imaging differentiates multiple system atrophy from Parkinson's disease: a positron emission tomography study with [11C] PBR28 and machine learning analysis. *Mov Disord* 37:119–129
31. Khawaldeh S, Tinkhauser G, Shah SA et al (2020) Subthalamic nucleus activity dynamics and limb movement prediction in Parkinson's disease. *Brain* 143:582–596
32. Poston KL, YorkWilliams S, Zhang K et al (2016) Compensatory neural mechanisms in cognitively unimpaired Parkinson disease. *Ann Neurol* 79:448–463
33. Trezzi J-P, Galozzi S, Jaeger C et al (2017) Distinct metabolomic signature in cerebrospinal fluid in early parkinson's disease. *Mov Disord* 32:1401–1408
34. Vanneste S, Song J-J, De Ridder D (2018) Thalamocortical dysrhythmia detected by machine learning. *Nat Commun* 9:1103
35. Goetz CG, Fahn S, Martinez-Martin P et al (2007) Movement Disorder Society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): process, format, and clinimetric testing plan. *Mov Disord* 22:41–47
36. Evers LJW, Krijthe JH, Meinders MJ et al (2019) Measuring Parkinson's disease

- over time: the real-world within-subject reliability of the MDS-UPDRS. *Mov Disord* 34: 1480–1487
37. Ahlrichs C, Samà A, Lawo M et al (2016) Detecting freezing of gait with a tri-axial accelerometer in Parkinson's disease patients. *Med Biol Eng Comput* 54:223–233
 38. Aich S, Pradhan PM, Park J et al (2018) A validation study of freezing of gait (FoG) detection and machine-learning-based FoG prediction using estimated gait characteristics with a wearable accelerometer. *Sensors* 18: 3287
 39. Borzì L, Mazzetta I, Zampogna A et al (2021) Prediction of freezing of gait in Parkinson's disease using wearables and machine learning. *Sensors* 21:614
 40. Dvorani A, Waldheim V, Jochner MCE et al (2021) Real-time detection of freezing motions in Parkinson's patients for adaptive gait phase synchronous cueing. *Front Neurol* 12:720516
 41. Shalin G, Pardoe S, Lemaire ED et al (2021) Prediction and detection of freezing of gait in Parkinson's disease from plantar pressure data using long short-term memory neural-networks. *J Neuroeng Rehabil* 18:1–15
 42. Park KW, Lee E-J, Lee JS et al (2021) Machine learning-based automatic rating for cardinal symptoms of Parkinson disease. *Neurology* 96:e1761–e1769
 43. Kim HB, Lee WW, Kim A et al (2018) Wrist sensor-based tremor severity quantification in Parkinson's disease using convolutional neural network. *Comput Biol Med* 95:140–146
 44. Eskofier BM, Lee SI, Daneault J-F, et al (2016) Recent machine learning advancements in sensor-based mobility analysis: deep learning for Parkinson's disease assessment. In: 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp 655–658
 45. Litvan I, Aarsland D, Adler CH et al (2011) MDS task force on mild cognitive impairment in Parkinson's disease: critical review of PD-MCI. *Mov Disord* 26:1814–1824
 46. Abós A, Baggio HC, Segura B et al (2017) Discriminating cognitive status in Parkinson's disease through functional connectomics and machine learning. *Sci Rep* 7:45347
 47. Betrouni N, Delval A, Chaton L et al (2019) Electroencephalography-based machine learning for cognitive profiling in Parkinson's disease: preliminary results. *Mov Disord* 34: 210–217
 48. García AM, Arias-Vergara TC, Vasquez-Correa J et al (2021) Cognitive determinants of dysarthria in Parkinson's disease: an automated machine learning approach. *Mov Disord* 36: 2862–2873
 49. Morales DA, Vives-Gilabert Y, Gómez-Ansón B et al (2013) Predicting dementia development in Parkinson's disease using Bayesian network classifiers. *Psychiatry Res* 213:92–98
 50. Shibata H, Uchida Y, Inui S et al (2022) Machine learning trained with quantitative susceptibility mapping to detect mild cognitive impairment in Parkinson's disease. *Parkinsonism Relat Disord* 94:104–110
 51. Hannink J, Kautz T, Pasluosta CF et al (2017) Sensor-based gait parameter extraction with deep convolutional neural networks. *IEEE J Biomed Health Inform* 21:85–93
 52. Lu M, Zhao Q, Poston KL et al (2021) Quantifying Parkinson's disease motor severity under uncertainty using MDS-UPDRS videos. *Med Image Anal* 73:102179
 53. Gao C, Sun H, Wang T et al (2018) Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease. *Sci Rep* 8:7129
 54. Severson KA, Chahine LM, Smolensky LA et al (2021) Discovery of Parkinson's disease states and disease progression modelling: a longitudinal data study using machine learning. *Lancet Digit Health* 3:e555–e564
 55. Salmanpour MR, Shamsaei M, Hajianfar G et al (2022) Longitudinal clustering analysis and prediction of Parkinson's disease progression using radiomics and hybrid machine learning. *Quant Imaging Med Surg* 12:90619–90919
 56. Oxtoby NP, Leyland L-A, Aksman LM et al (2021) Sequence of clinical and neurodegeneration events in Parkinson's disease progression. *Brain J Neurol* 144:975–988
 57. Latourelle JC, Beste MT, Hadzi TC et al (2017) Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed Parkinson's disease: a longitudinal cohort study and validation. *Lancet Neurol* 16:908–916
 58. Ahmadi Rastegar D, Ho N, Halliday GM et al (2019) Parkinson's progression prediction using machine learning and serum cytokines. *NPJ Park Dis* 5:1–8
 59. Amara AW, Chahine LM, Caspell-Garcia C et al (2017) Longitudinal assessment of excessive daytime sleepiness in early Parkinson's disease. *J Neurol Neurosurg Psychiatry* 88:653–662
 60. Couronne R, Vidailhet M, Corvol JC, et al (2019) Learning disease progression models with longitudinal data and missing values. In: 2019 IEEE 16th international symposium on

- biomedical imaging (ISBI 2019), pp 1033–1037
61. Faouzi J, Bekadar S, Artaud F et al (2022) Machine learning-based prediction of impulse control disorders in Parkinson's disease from clinical and genetic data. *IEEE Open J Eng Med Biol* 3:96–107
 62. You H, Mariani L-L, Mangone G et al (2018) Molecular basis of dopamine replacement therapy and its side effects in Parkinson's disease. *Cell Tissue Res* 373:111–135
 63. Yang B, Wang X, Mo J et al (2021) The amplitude of low-frequency fluctuation predicts levodopa treatment response in patients with Parkinson's disease. *Parkinsonism Relat Disord* 92:26–32
 64. Kim Y, Suescun J, Schiess MC et al (2021) Computational medication regimen for Parkinson's disease using reinforcement learning. *Sci Rep* 11:9313
 65. Perestelo-Pérez L, Rivero-Santana A, Pérez-Ramos J et al (2014) Deep brain stimulation in Parkinson's disease: meta-analysis of randomized controlled trials. *J Neurol* 261: 2051–2060
 66. Boutet A, Madhavan R, Elias GJB et al (2021) Predicting optimal deep brain stimulation parameters for Parkinson's disease using functional MRI and machine learning. *Nat Commun* 12: 3043
 67. Geraedts VJ, Koch M, Contarino MF et al (2021) Machine learning for automated EEG-based biomarkers of cognitive impairment during deep brain stimulation screening in patients with Parkinson's disease. *Clin Neurophysiol* 132:1041–1048
 68. Phokaewvarangkul O, Vateekul P, Wichakam I et al (2021) Using machine learning for predicting the best outcomes with electrical muscle stimulation for tremors in Parkinson's disease. *Front Aging Neurosci* 13:727654
 69. Panyakaew P, Pornputtapong N, Bhidayasiri R (2021) Using machine learning-based analytics of daily activities to identify modifiable risk factors for falling in Parkinson's disease. *Parkinsonism Relat Disord* 82:77–83

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 27

Machine Learning in Neuroimaging of Epilepsy

**Hyo Min Lee, Ravnoor Singh Gill, Neda Bernasconi,
and Andrea Bernasconi**

Abstract

Epilepsy is a prevalent chronic condition affecting about 50 million people worldwide. A third of patients suffer from seizures unresponsive to medication. Uncontrolled seizures damage the brain, are associated with cognitive decline, and have negative impact on well-being. For these patients, the surgical resection of the brain region that gives rise to seizures is the most effective treatment. In this context, due to its unmatched spatial resolution and whole-brain coverage, magnetic resonance imaging (MRI) plays a central role in detecting lesions. The last decade has witnessed an increasing use of machine learning applied to multimodal MRI, which has allowed the design of tools for computer-aided diagnosis and prognosis. In this chapter, we focus on automated algorithms for the detection of epileptogenic lesions and imaging-derived prognostic markers, including response to anti-seizure medication, postsurgical seizure outcome, and cognitive reserves. We also highlight advantages and limitations of these approaches and discuss future directions toward person-centered care.

Key words Epilepsy, Focal cortical dysplasia, Temporal lobe epilepsy

1 Introduction

Epilepsy is a prevalent chronic condition affecting about 50 million people worldwide. Seizures are generally defined as transient symptoms and signs due to excessive neuronal activity; based on these manifestations, they can be classified as focal or generalized. Various etiologies have been associated with epilepsy, including structural, genetic, infectious, metabolic, and immune. Frequent structural pathologies include traumatic brain injury, tumors, vascular malformations, stroke, and developmental disorders. A third of patients suffer from seizures unresponsive to medication [1]. Drug-resistant seizures damage the brain [2] and are associated with high risks for socioeconomic difficulties, cognitive decline, and mortality [3]. The main forms of drug-resistant focal epilepsy are related to focal cortical dysplasia (FCD), a structural brain developmental malformation, and mesiotemporal lobe sclerosis, a

histopathological lesion that combines various degrees of neuronal loss and gliosis in the hippocampus and adjacent cortices. To date, the most effective treatment has been the surgical resection of these structural lesions. In this context, magnetic resonance imaging (MRI) has been instrumental in the pre-surgical evaluation, as it can reliably detect these anomalies due to its unmatched spatial resolution and whole-brain coverage. Indeed, localizing a structural lesion on MRI is the strongest predictor of favorable seizure outcome after surgery [4–6]. Yet, challenges remain. Large numbers of patients have subtle lesions undetected on routine MRI. In these patients, referred to as “MRI-negative,” the surgical outcome is poorer compared to those in whom a structural lesion is identified [7]. Moreover, even in carefully selected patients, about 30% may continue having seizures after surgery. These shortcomings have motivated the development of advanced analytic techniques for the discovery of diagnostic and prognostic biomarkers, which serve as input to machine learning. MRI quantitation holds promise to match or exceed the evaluation by human experts. In this chapter, we will describe algorithms for the detection of epileptogenic lesions, prediction of clinical outcomes, and identification of disease subtypes in drug-resistant focal epilepsy. We will highlight their advantages and limitations and discuss future directions toward personalized care.

2 Lesion Mapping

In epilepsy, identifying a structural lesion on MRI is crucial for successful surgery [5]. Advances in MRI acquisition technology, specifically high (3T) and ultrahigh (7T) field imaging combined with multiple phased array head coils, have permitted precise lesion characterization. Machine learning holds great promise for exceeding human performance [8]. Indeed, application on structural MRI data has enabled increasingly reliable detection of epileptogenic lesions, including those overlooked on routine radiological examination. Automated lesion detection is generally performed by supervised classifiers that are trained to learn the distributions and inter-relations between MRI features that distinguish lesional from non-lesional tissue, leveraging this knowledge to classify a given tissue type in previously unseen patients.

2.1 Mapping *Hippocampal Sclerosis in Temporal Lobe Epilepsy*

Temporal lobe epilepsy (TLE), the most common focal syndrome in adults, is pathologically defined by varying degrees of neuronal loss and gliosis in the hippocampus and adjacent structures [9]. On MRI, marked hippocampal sclerosis (HS) appears as atrophy and signal hyperintensity, generally more severe ipsilateral to the seizure focus. Accurate identification of hippocampal atrophy as a marker of HS is crucial for deciding the side of surgery. While volumetry

has been one of the first computational analyses applied to TLE [10–15], the need for accurate localization of pathology has motivated a move from whole-structure volumetry to surface-based approaches allowing a precise mapping of anomalies along the hippocampal axis. In this context, 3D surface-based shape models permit localizing regional morphological differences that may not be readily identifiable [16]. Surface modeling based on spherical harmonics [17] has been particularly performant [18]. Following this method, hippocampal labels are processed using a series of spherical harmonics with increasing degree of complexity to parametrize their surface boundary. Anatomical intersubject correspondence is guaranteed by aligning the surfaces of each individual to the centroid and the longitudinal axis of the first-order ellipsoid of the mean surface template derived from controls and patients. Computing the Jacobian determinants of the surface displacement vectors allows quantifying localized areas of atrophy [18, 19]. Overall, surface-based methods have proven superior to their volumetric counterparts not only in terms of segmentation performance [20] but also in predicting clinical outcomes as well as mapping disease progression [21, 22]. Applying clustering to surface-based morphometry of the hippocampus, amygdala, and entorhinal cortex, a clinically homogeneous cohort of drug-resistant TLE patients with a unilateral seizure focus could be segregated into classes with distinct MRI and histopathological signatures [23]. Extending this methodology by extracting features along the medial surface of hippocampal subfields has allowed to further probe the laminar integrity of this structure [24, 25].

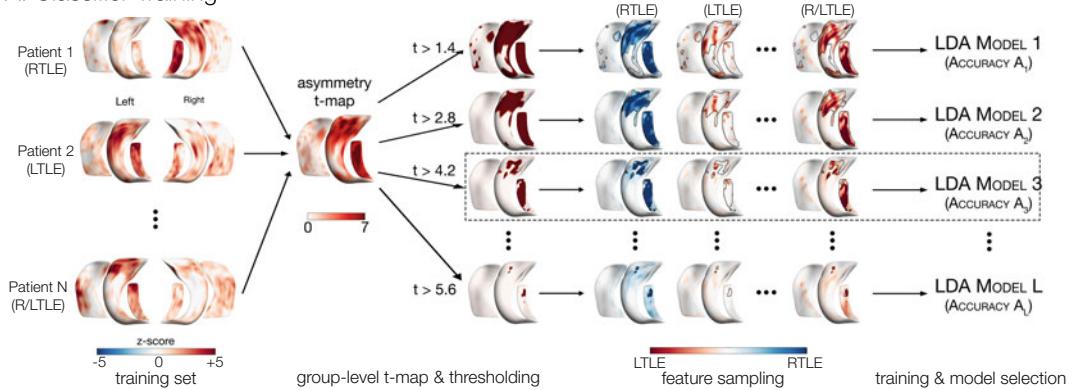
Manual hippocampal volumetry is time-prohibitive and prone to rater bias. These challenges, together with increasing demand to study larger patient cohorts, have motivated the shift toward automated segmentation, setting the basis for large-scale clinical use. Initial methods for whole hippocampal segmentation used a single template or deformable models constrained by shape priors obtained from neurotypical individuals [26–29]. More recent approaches rely on multiple templates and label fusion; by selecting a subset of atlases from a template library which best fit the structure to segment, thereby accounting for intersubject variability, these approaches have provided increased performance [30–32]. In epilepsy, SurfMulti achieved identical performance in TLE (Dice: 86.9%) and healthy controls (87.5%), outperforming the widely used FreeSurfer, even in the presence of prevalent atypical hippocampal morphology (i.e., maldevelopment or malrotation) and significant atrophy [20]. Advances in MR acquisition hardware and sequence technology, which enable submillimetric resolution and improved signal-to-noise ratio, have facilitated accurate identification of hippocampal subfields or subregions, including the dentate gyrus, subiculum, and the cornu ammonis (CA1–4) regions

[33]. Several methods have been developed for MRI-based subfield segmentation [19, 34–38], providing an average Dice of 88%, with fast inference times. Among them, the SurfPatch subfield segmentation algorithm, operating on T1-weighted MRI, combines multiple templates, parametric surfaces, and patch-based sampling for compact representation of shape, texture, and intensity [38]. SurfPatch showed high segmentation accuracy (Dice >0.82 for all subfields) and robustness to the size of template library and image resolution (millimetric and sub-millimetric) while demonstrating utility for reliable TLE lateralization (93% accuracy).

Brain segmentation may serve as the basis to extract features used to train classifiers for predictions. An SVM-based classifier using volumetric features derived from whole-brain T1-weighted images was able to classify and lateralize TLE [39]. However, regions identifying TLE groups were primarily located outside the mesiotemporal lobe, making such design impractical for previously unseen cases and difficult to interpret in MRI-negative patients. Overall, while high lateralization performance (>90%) may be achieved in MRI-positive patients, the yield in MRI-negative TLE remains at less than 20% when using features derived from T1-weighted images [40, 41]. On the other hand, classifiers operating on FLAIR [42] and double inversion recovery [43] have shown 70% lateralization in MRI-negative patients. Yet, studies have been rather limited in sample size and have lacked histological verification or long-term measures of seizure outcome after surgery; moreover, absence of validation in independent datasets has precluded assessment of generalizability. To tackle these shortcomings, our group recently designed an automated surface-based linear discriminant classifier trained on T1- and FLAIR-derived laminar features of HS (Fig. 1) [44]. As HS is typically characterized by T1-weighted hypointensity and T2-weighted hyperintensity, the synthetic contrast FLAIR/T1 maximized their combined contributions to detect the full pathology spectrum. The classifier accurately lateralized the focus in 85% of patients with MRI-negative but histologically verified HS. Notably, similar high performance was achieved in two independent validation cohorts, thereby establishing generalizability across cohorts, scanners, and parameters. Such validated classifiers set the basis for broad clinical translation.

Recently, the widespread adoption of deep learning in medical imaging has promoted a resurgence in volumetric segmentation methods. Unlike contemporary algorithms, deep learning does not require the data to be extensively preprocessed, thus eliminating the need to build template libraries. More specifically, the ability of convolutional neural networks to learn salient features from multimodal data in the course of the training process rather than using hand-crafted features has enabled them to outperform

A. Classifier Training



B. Representative MRI-negative Patient

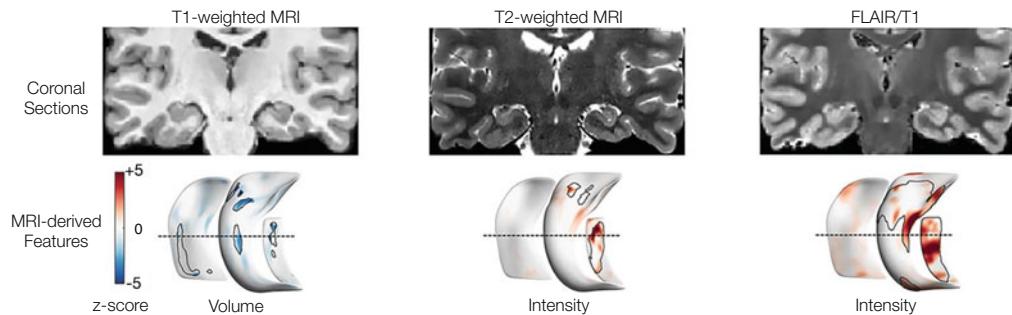


Fig. 1 Automated lateralization of hippocampal sclerosis. **(a)** In the training phase, an optimal region of interest is defined for each modality to systematically sample features (T1-derived volume, T2-weighted intensity, and FLAIR/T1 intensity) across individuals. To this purpose, in each patient paired t-tests compare corresponding vertices of the left and right subfields, z-scored with respect to healthy controls. The resulting group-level asymmetry t-map is then thresholded from 0 to the highest value and binarized; for each threshold, the binarized t-map is overlaid on the asymmetry map of each individual to compute the average across subfields. Then a linear discriminant classifier is trained for each threshold, and the model yielding the highest lateralization accuracy (in this example LDA model 3) is used to test the classifier. **(b)** Lateralization prediction in a patient with MRI-negative left TLE. Coronal sections are shown together with the automatically generated asymmetry maps for columnar volume, T2-weighted, and FLAIR/T1 intensities. On each map, dotted line corresponds to the level of the coronal MRI section and the optimal ROI obtained during training is outlined in black

traditional approaches, with Dice overlap indices exceeding 90% in both healthy [44–46] and atrophic [47] hippocampi. Deep learning applications for seizure focus lateralization have insofar been limited. One study showed that deep learning classifiers performed similar or worse than SVM-based classifiers [48]; this work, however, explored only a singular set of hyperparameters using pre-defined features for the neural networks, thereby missing the opportunity to exploit hierarchical feature learning, one of the most distinctive characteristics of deep learning.

2.2 Automated Detection of Focal Cortical Dysplasia

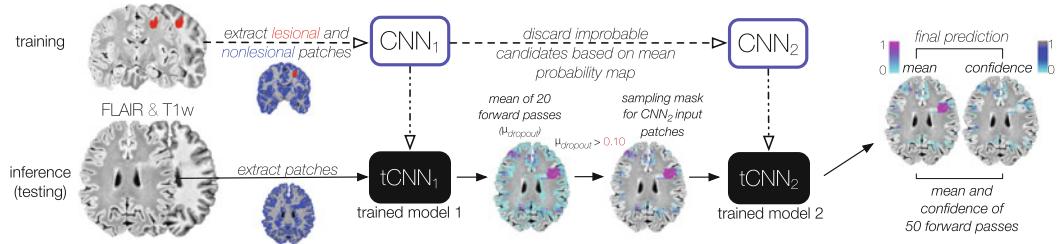
On MRI, focal cortical dysplasia (FCD) presents with a visibility spectrum encompassing variable degrees of gray matter (GM) and white matter (WM) changes that can challenge visual identification. Indeed, recent series indicate that up to 33% of FCD Type II, the most common surgically amenable developmental malformation, present with “unremarkable” routine MRI, even though typical features are ultimately identified in the histopathology of the resected tissue [49–51]. These so-called “MRI-negative” FCDs represent a major diagnostic challenge. Indeed, to define the epileptogenic area, patients undergo long and costly hospitalizations for EEG monitoring with intracerebral electrodes, a procedure that carries risks similar to surgery itself [52, 53]. Moreover, patients without MRI evidence for FCD are less likely to undergo surgery and consistently show worse seizure control compared to those with visible lesions [4, 54, 55]. This clinical difficulty has motivated the development of computer-aided methods aimed at optimizing detection *in vivo*. Such techniques provide distinct information through quantitative assessment without the cost of additional scanning time.

Early methods opted for voxel-based methods to quantify group-level structural abnormalities related to MRI-visible dysplasias by thresholding GM concentration (e.g., >1 SD relative to the mean in healthy controls). While such methods are sensitive (87–100%) in detecting conspicuous malformations, they fail to identify two-thirds of subtle, MRI-negative lesions [56–59]. To counter the relative lack of specificity, our group introduced an original approach to integrate key voxel-wise textures and morphological modeling (i.e., cortical thickening, blurring of the GM-WM junction, and intensity alterations) derived from T1-weighted images into a composite map [60, 61]. The clinical value of this computer-aided visual identification was supported by its 88% sensitivity and 95% specificity, vastly outperforming conventional MRI. An alternative method quantifies blurring as voxels that belong neither to GM or WM [62]. Integrating morphological operators with higher-order image texture features invisible to the human eye into a fully automated classifier provided a sensitivity of 80% [63, 64]. In contrast to voxel-based methods, surface-based morphometry offers an anatomically plausible quantification of structural integrity that preserves cortical topology. Surface-based modeling of cortical thickness, folding complexity, and sulcal depth, together with intra- and subcortical mapping of MRI intensities and textures, allow for a more sensitive description of FCD pathology. Over the last decade, several such algorithms have been developed, with detection rates up to 83% [65–71]. The addition of FLAIR has contributed to further increase in sensitivity, particularly for the detection of smaller lesions [66]. Notably, an integration of surface-based methods into clinical workflow would be contingent to careful verification of preprocessing steps, including manual

corrections of tissue segmentation and surface extraction to obtain high-fidelity FCD features. Without such careful and intensive data preprocessing and inspection, the performance is rather poor, as demonstrated by a recent multicenter study in which the sensitivity was below 70% with a specificity close to chance level even in MRI visible lesions [72].

Despite efforts dedicated to the development of increasingly sophisticated detection algorithms, some pitfalls are to be considered. Algorithms have not been systematically validated with histologically verified lesions or independent datasets. Many have not been tested or fail in MRI-negative cases. In general, detection algorithms have assumed structural anomalies to be homogeneous across lesions and patients, a notion challenged by recent histopathological [73, 74] and genetic [75] data. Moreover, they rely on limited number of features designed by human experts based on their knowledge, which may not capture the full pathological complexity. Importantly, the deterministic nature of these algorithms does not permit risk assessment, a necessity for integration into clinical diagnostic systems. Currently, benchmark automated detection fails in 20–40% of patients, particularly those with subtle FCD, and suffers from high false-positive rates. Relative to conventional methods, in recent years, deep neural networks have shown high sensitivity at detection across various diseases [see 76, 77, for review]. Specifically, convolutional neural networks learn abstract concepts from high-dimensional data alleviating the challenging task of handcrafting features [78]. To date, a few studies have used deep learning for FCD detection [79–81]. However, their clinical description has been scarce or absent, and the information on how lesions were labeled for the training as well as histological validation was not provided. Notably, while their performance was reasonably high in MRI-positive cohorts (range: 85–92%; no MRI-negative cases identified) using either T1-weighted or T2-weighted FLAIR images, sample sizes were limited to 10–40 and sourced from a single center. Deep learning requires large corpus of expertly labeled annotations (ground truth) to train and optimize the network, both cost- and time-consuming endeavors, resulting in suboptimal cohort sizes. To overcome this challenge, our group leveraged a patch-based augmentation that extracts several hundreds of overlapping patches from a single subject, thereby scaling up the data without the requirement of an impractically large cohort [82]. This deep learning algorithm relied on clinically available T1- and T2-weighted FLAIR MRI of a large cohort of patients with histologically validated lesions, collated across multiple tertiary epilepsy centers (Fig. 2). Notably, operating on 3D voxel space (i.e., in true volumetric domain) allowed assessing the spatial neighborhood of the lesion, whereas prior surface-based methods have considered each vertex location independently. This convolutional neural network classifier yields the highest

A. Classifier Design



B. Detection and Confidence

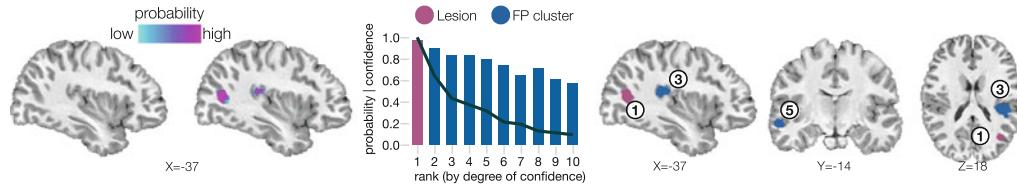


Fig. 2 Automated FCD detection using deep learning. **(a)** The training and testing workflow. In this cascaded system, the output of the convolutional neural network 1 (CNN_1) serves as an input to CNN_2 . CNN_1 maximizes the detection of lesional voxels; CNN_2 reduces the number of misclassified voxels, removing false positives (FPs) while maintaining optimal sensitivity. The training procedure (dashed arrows) operating on T1-weighted and FLAIR extracts 3D patches from lesional and non-lesional tissue to yield $t\text{CNN}_1$ (trained model 1) and $t\text{CNN}_2$ with optimized weights (vertical dashed-dotted arrows). These models are then used for subject-level inference. For each unseen subject, the inference pipeline (solid arrows) uses $t\text{CNN}_1$ and generates a mean (μ_{dropout}) of 20 predictions (forward passes); the mean map is then thresholded voxel-wise to discard improbable lesion candidates ($\mu_{\text{dropout}} > 0.1$). The resulting binary mask serves to sample the input patches for the $t\text{CNN}_2$. A mean probability and uncertainty maps are obtained by collating 50 predictions; uncertainty is transformed into confidence. The sampling strategy (identical for training and inference) is only illustrated for testing. **(b)** Sagittal sections show the native T1-weighted MRI superimposed with the lesion probability map. The bar plot shows the probability of the lesion (purple) and false-positive (FP, blue) clusters sorted by their rank; the superimposed line indicates the degree of confidence for each cluster. In this example, the lesion (cluster 1 in purple) has both the highest probability and confidence

performance to date with a sensitivity of 93% using a leave-one-site-out cross-validation and 83% when tested on an independent cohort while maintaining a high specificity of 89% both in healthy and disease controls. Importantly, deep learning detected MRI-negative FCD with 85% sensitivity, thus offering a considerable gain over standard radiological assessment. Results were generalizable across cohorts with variable age, hardware, and sequence parameters. Using Bayesian uncertainty estimation that enables risk stratification [83, 84], our predictions were stratified according to the confidence to be truly lesional. In 73% of cases, the FCD was among the top five clusters with the highest confidence to be lesional; in half of them, it ranked the highest. Ranking putative lesional clusters in each patient based on confidence helps the examiner to gauge the significance of all findings. In other words,

by pairing predictions with risk stratification, this classifier may assist clinicians to adjust hypotheses relative to other tests, thus increasing diagnostic confidence. Taken together, such characteristics and performance promise great potential for broad clinical translation.

3 Prediction of Clinical Outcomes

While science investigating the neurobiology of epilepsy has been growing rapidly, translating knowledge into clinical practice has been limited. Specifically, individualized predictions of drug resistance, surgical outcome, and cognitive dysfunction have been attempted with limited success [85]. For example, early investigations that aimed to predict anti-seizure medication response used machine learning on genomic data (viz., single nucleotide polymorphisms) and showed limited generalizability with inconsistent performance across studies [86–88]. Similarly, other models trained on electro-clinical and demographic features of thousands of patients [89–92] achieved high sensitivity (>90%) but unacceptably low specificity (<25%). Importantly, no external validation was performed on independent cohorts. The prediction of seizure outcome after surgery has been extensively explored in TLE patients. Some of the early investigations relied on clinical [93] and neuropsychological features [94], achieving high performance, but in limited samples of less than 20 patients. Given the increasing conceptualization of TLE as a system-level disorder, numerous studies have tested the hypothesis that structural and functional alterations beyond the mesial temporal lobe may contribute to negative seizure outcome [95, 96]. For instance, WM microstructural features derived from diffusion tensor imaging have shown to achieve high sensitivity (70–86%) but modest specificity (65–70%) [97, 98]. Other studies have relied on connectivity features for prediction; these include nodal hubness of the thalamus and whole-brain distance-based measures of functional connectivity, which achieve an accuracy at about 75% but modest specificity (ranging from 35 to 62%) [99, 100]. Conversely, while topological features of structural connectome have generally shown high predictive value for favorable postsurgical outcome, with an area under the receiver operating characteristics of 0.88, specificity for prediction of seizure relapse is low (29–54%) [101, 102]. Overall, the lack of large-scale external validation and relatively low specificity of these models need to be addressed to establish their generalizability and potential clinical use.

3.1 Disease Biotyping: Leveraging Individual Variability to Optimize Predictions

To date, most neuroimaging studies of epilepsy have been based on “one-size-fits-all” group-level analytical approaches. While such study designs can isolate reliable and consistent average group-level differences, they merely decipher the common patterns without modeling the inter-individual variations along the disease spectrum [103]. Conversely, the conceptualization of epilepsy as a heterogeneous disorder and explicit modeling of inter-individual phenotypic variations may be exploited to predict individual-specific clinical outcomes [104].

Over the past decades, FCD characterization has been driven by histology, with the primary objective to establish subtype-specific imaging signatures [105]. Although histological grading is a well-defined framework, the current approach is based on descriptive criteria that do not consider the severity of each feature, thereby limiting neurobiological understanding. The ability to perform *in vivo* patient stratification is gaining relevance due to the emergence of minimally invasive surgical procedures that do not provide specimens for histological examination [106]. From a neurobiological standpoint, whether FCD IIB (dysmorphic neurons and balloon cells) and IIA (dysmorphic neurons only) subtypes represent etiologically distinct entities, or a spectrum is a matter of debate. Recent studies have shown significant cellular variability, with anomalies that may vary across lesions within the same subtype [73]. Moreover, multiple subtypes may coexist within the same FCD, with the most severe phenotype determining the final diagnosis [74]. Furthermore, recent studies have identified regulatory genes of the mTOR pathway that cause FCD via somatic mutations, revealing a genetic continuum not linked to discrete FCD subtypes [75]. Hence, assessing the intra- and inter-lesional variability on MRI may offer a novel basis to advance our understanding of FCD neurobiology and improve lesion detection. Leveraging hierarchical clustering to model connectivity from FCD tissue to the rest of the cortex demonstrated that network dysfunction can dissociate patients with excellent from those with suboptimal post-surgical seizure outcomes [107]. Another recent work applying consensus clustering to multi-contrast 3T MRI uncovered FCD tissue classes with distinct structural profiles, variably expressed within and across patients [108]. Importantly, these classes had differential histopathological embeddings, and their clinical utility was supported by gain in performance of a lesion detection algorithm trained on class-informed data compared to class-naïve paradigm.

In TLE, histopathological reports have shown substantial variability in the distribution and severity of mesiotemporal lobe sclerosis between patients [109, 110]. A modern approach combining quantitative histology and unsupervised machine learning identified histological subtypes with differential severity and regional signatures [111]. Motivated by these findings, recent studies have

exploited inter-individual variability of imaging or cognitive phenotypes to optimize predictions of clinical outcomes. The first attempts were based on categorical models, which provided subtypes of patients with a given phenotype. Clustering applied to surface-based morphometry uncovered four TLE subtypes having distinct subregional patterns of mesiotemporal atrophy [23]. These four subtypes differed with respect to histopathology and postsurgical seizure outcome. Classifiers operating on class membership accurately predicted surgical outcome in >90% of patients, outperforming learners trained on conventional MRI volumetry. In the context of cognition, unsupervised techniques have identified phenotypes, such as language and memory impairment associated with distinct patterns of WM microstructural damage [112] and connectome disorganization [113].

Compared to categorical models such as clustering, dimensional approaches allow a more in-depth conceptualization of inter-individual variability by uncovering axes of pathology that are co-expressed within and between individuals. In other words, such approaches allow patients to express multiple disease factors to varying degrees rather than assigning subjects to a single subtype. Applying latent Dirichlet allocation, an unsupervised technique derived from topic modeling, to multimodal MRI features of hippocampal and whole-brain GM and WM pathology, a recent study uncovered dimensions of heterogeneity (or disease factors) in TLE that were not expressed in healthy controls and only minimally in patients with frontal lobe epilepsy, supporting specificity (Figs. 3 and 4) [114]. Importantly, classifiers trained on the patients' factor composition predicted response to anti-seizure medications (76% accuracy) and surgery (88%) as well as cognitive scores for verbal IQ, memory, and sequential motor tapping, outperforming learners trained on group-level data [114]. In translational terms, assessing inter-individual variability through dimensional modeling mines clinically relevant disease characteristics that would otherwise be missed.

4 Conclusion and Future Perspectives

Machine learning applied to MRI has successfully uncovered mesoscopic structural and functional biomarkers predictive of clinical outcomes. Overall, the most significant impact has been the development of lesion detection algorithms that have transformed MRI-negative into MRI-positive, thus offering the life-changing benefits of epilepsy surgery to more patients. More recently, biotyping techniques exploiting intra- and intersubject variability have permitted to further optimize the prediction of outcomes. Integrating such approaches with other domains such as genomics

A. Latent Factor Analysis

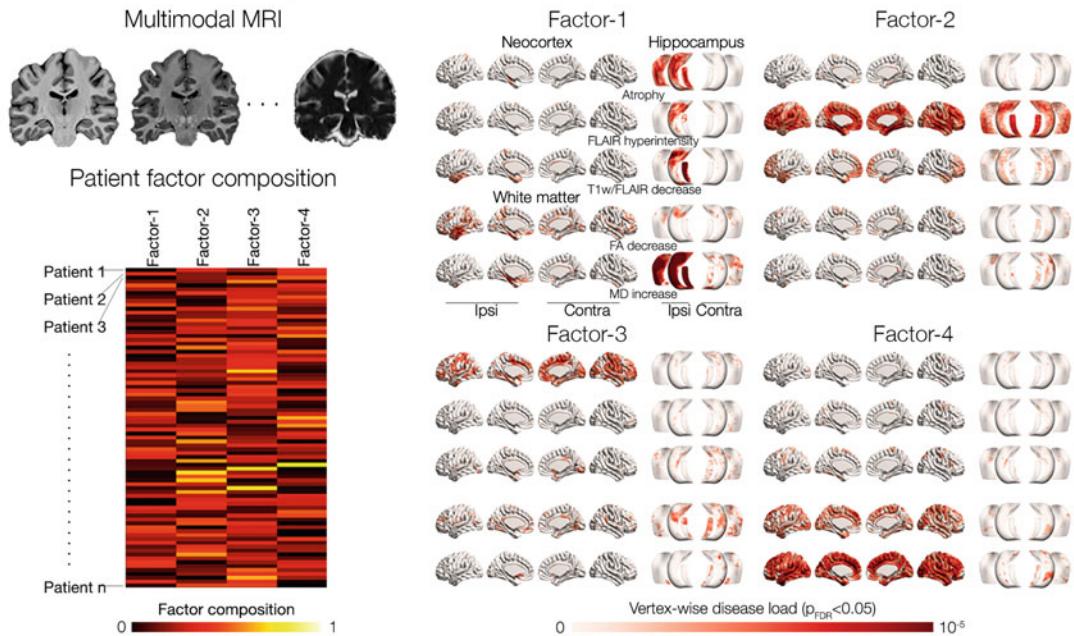


Fig. 3 Latent disease factors in TLE. Multimodal MRI (T1w, FLAIR, T1w/FLAIR, diffusion-derived FA, and MD) is combined with surface-based analysis to model the main features of TLE pathology (atrophy, gliosis, demyelination, and microstructural damage), which are z-scored with respect to the analogous vertices of healthy controls' ipsi- and contralateral to the seizure focus. Latent Dirichlet allocation uncovered four latent relations (viz., disease factors) from these features (expressed as posterior probability) and quantified their co-expression (ranging from 0 to 1) as shown in the patients' factor composition matrix. On the color scale below, the disease factor maps higher probability (darker red) and signifies a greater contribution of a given feature to the factor, namely, the disease load ($p_{FDR} < 0.05$)

promises to elucidate molecular mechanisms that drive MRI phenotypes, offering novel avenues to study disease processes [115, 116].

Notwithstanding its diagnostic capabilities, machine learning is still viewed by some as a “black box,” possibly due to the increasing complexity of the predictive models, particularly those relying on deep learning [117]. In this regard, increased model interpretability may prevent biases and reduce the risk of incorrect clinical inferences. It is, therefore, crucial to understand how the model arrived at a particular decision. For large-scale neural networks, this may be achieved by visualizing on a map the features learned in the course of training. Besides transparency, significant obstacles to clinical adoption are privacy and ethics. These concerns have been circumvented so far through single site designs or multi-institutional training aggregating data in a single center. While the latter allows addressing model generalizability through physical access to independent datasets, federated learning may provide decentralized collaborations without data sharing [30]. As the

B. Individualized Predictions

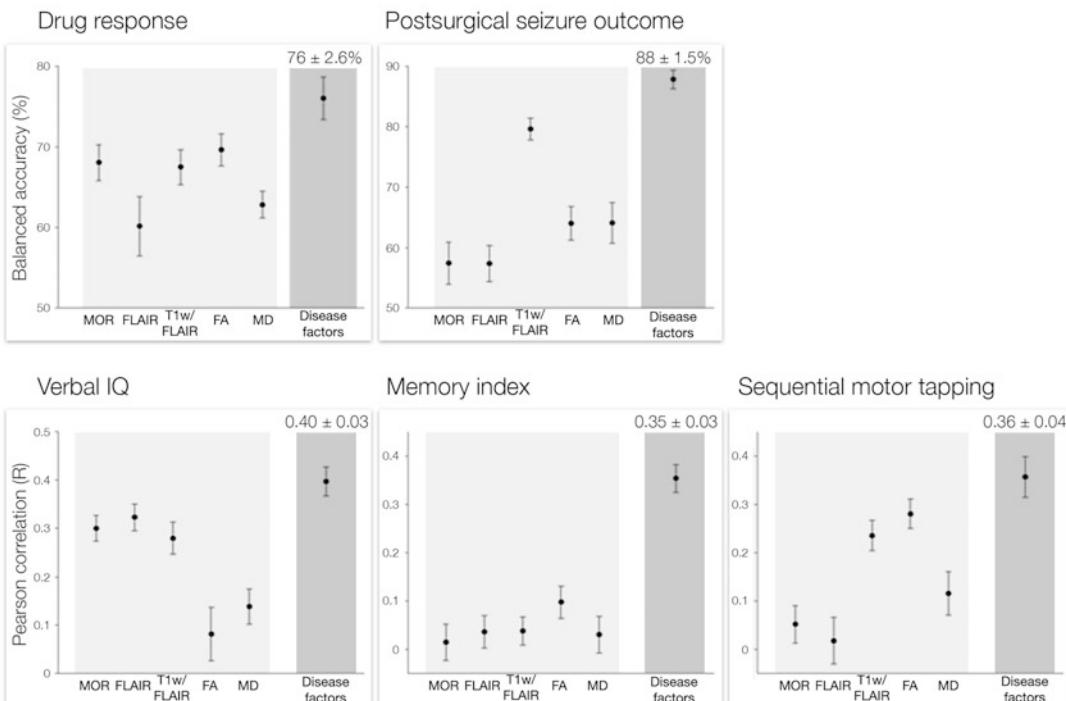


Fig. 4 Latent disease factors in TLE. Drug response, seizure outcome, verbal IQ, memory index, and motor index are more accurately predicted when using latent disease factors than when relying on conventional group-level features ($p\text{FDR} < 0.001$). Data points indicate mean balanced accuracy for categorical data (drug-response, seizure outcome) and Pearson correlation coefficients for numerical data (cognitive scores) evaluated based on 100 repetitions of tenfold cross-validation

data corpus diversifies and expands to include more edge cases, performance and confidence of future classifiers will inevitably improve. Ultimately, clinical translation of complex techniques into practice is contingent to continued efforts in education of clinicians combined with increased accessibility to source codes and algorithms.

References

- Wiebe S, Jette N (2012) Pharmacoresistance and the role of surgery in difficult to treat epilepsy. *Nat Rev Neurol* 8:669. <https://doi.org/10.1038/nrneurol.2012.181>
- Caciagli L, Bernasconi A, Wiebe S, Koepp MJ, Bernasconi N, Bernhardt BC (2017) A meta-analysis on progressive atrophy in intractable temporal lobe epilepsy. *Time is brain?* 89(5): 506–516. <https://doi.org/10.1212/wnl.000000000004176>
- Keezer MR, Sisodiya SM, Sander JW (2016) Comorbidities of epilepsy: current concepts and future perspectives. *Lancet Neurol* 15(1):106–115
- Jobst BC, Cascino GD (2015) Resective epilepsy surgery for drug-resistant focal epilepsy: a review. *JAMA* 313(3):285–293. <https://doi.org/10.1001/jama.2014.17426>
- Téllez-Zenteno JF, Ronquillo LH, Moien-Afshari F, Wiebe S (2010) Surgical outcomes

- in lesional and non-lesional epilepsy: a systematic review and meta-analysis. *Epilepsy Res* 89(2):310–318. <https://doi.org/10.1016/j.eplepsyres.2010.02.007>
6. West S, Nevitt SJ, Cotton J, Gandhi S, Weston J, Sudan A, Ramirez R, Newton R (2019) Surgery for epilepsy. Cochrane Database Syst Rev 6:CD010541
 7. Bernasconi A, Bernasconi N, Bernhardt BC, Schrader D (2011) Advances in MRI for cryptogenic epilepsies. *Nat Rev Neurol* 7:99. <https://doi.org/10.1038/nrneurol.2010.199>
 8. Rauschecker AM, Rudie JD, Xie L, Wang J, Duong MT, Botzolakis EJ, Kovalovich AM, Egan J, Cook TC, Bryan RN (2020) Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI. *Radiology* 295(3): 626–637
 9. Blümcke I, Thom M, Aronica E, Armstrong DD, Bartolomei F, Bernasconi A, Bernasconi N, Bien CG, Cendes F, Coras R, Cross JH, Jacques TS, Kahane P, Mathern GW, Miyata H, Moshé SL, Oz B, Özkarla C, Perucca E, Sisodiya S, Wiebe S, Spreafico R (2013) International consensus classification of hippocampal sclerosis in temporal lobe epilepsy: a Task Force report from the ILAE Commission on Diagnostic Methods. *Epilepsia* 54(7):1315–1329. <https://doi.org/doi:10.1111/epi.12220>
 10. Cascino GD, Jack CR Jr, Parisi JE, Sharbrough FW, Hirschorn KA, Meyer FB, Marsh WR, O'Brien PC (1991) Magnetic resonance imaging-based volume studies in temporal lobe epilepsy: pathological correlations. *Ann Neurol* 30(1):31–36
 11. Cendes F, Andermann F, Gloor P, Evans A, Jones-Gotman M, Watson C, Melanson D, Olivier A, Peters T, Lopes-Cendes I (1993) MRI volumetric measurement of amygdala and hippocampus in temporal lobe epilepsy. *Neurology* 43(4):719–719
 12. Watson C, Jack CR, Cendes F (1997) Volumetric magnetic resonance imaging: clinical applications and contributions to the understanding of temporal lobe epilepsy. *Arch Neurol* 54(12):1521–1531
 13. Bernasconi N, Bernasconi A, Caramanos Z, Antel S, Andermann F, Arnold DL (2003) Mesial temporal damage in temporal lobe epilepsy: a volumetric MRI study of the hippocampus, amygdala and parahippocampal region. *Brain* 126(2):462–469
 14. Bernasconi N, Bernasconi A, Andermann F, Dubeau F, Feindel W, Reutens D (1999) Entorhinal cortex in temporal lobe epilepsy. *Quantitative MRI Study* 52(9):1870–1870. <https://doi.org/10.1212/wnl.52.9.1870>
 15. Bernasconi N, Bernasconi A, Caramanos Z, Dubeau F, Richardson J, Andermann F, Arnold D (2001) Entorhinal cortex atrophy in epilepsy patients exhibiting normal hippocampal volumes. *Neurology* 56(10): 1335–1339. <https://doi.org/10.1212/wnl.56.10.1335>
 16. Hogan RE, Wang L, Bertrand ME, Willmore LJ, Bucholz RD, Nassif AS, Csernansky JG (2004) MRI-based high-dimensional hippocampal mapping in mesial temporal lobe epilepsy. *Brain* 127(8):1731–1740
 17. Styner M, Oguiz I, Xu S, Brechbühler C, Pantazis D, Levitt JJ, Shenton ME, Gerig G (2006) Framework for the statistical shape analysis of brain structures using SPHARM-PDM. *Insight J* 1071:242
 18. Kim H, Besson P, Colliot O, Bernasconi A, Bernasconi N (2008) Surface-based vector analysis using heat equation interpolation: a new approach to quantify local hippocampal volume changes. In: Medical image computing and computer-assisted intervention—MICCAI 2008. Lecture Notes in Computer Science, vol 5241, pp 1008–1015. https://doi.org/10.1007/978-3-540-85988-8_120
 19. Kim H, Bernhardt BC, Kulaga-Yoskovitz J, Caldairo B, Bernasconi A, Bernasconi N (2014) Multivariate hippocampal subfield analysis of local MRI intensity and volume: application to temporal lobe epilepsy. In: Medical image computing and computer-assisted intervention—MICCAI 2014. Lecture Notes in Computer Science, vol 8674, pp 170–178
 20. Kim H, Mansi T, Bernasconi N, Bernasconi A (2012) Surface-based multi-template automated hippocampal segmentation: application to temporal lobe epilepsy. *Med Image Anal* 16(7):1445–1455. <https://doi.org/10.1016/j.media.2012.04.008>
 21. Bernhardt BC, Worsley KJ, Kim H, Evans AC, Bernasconi A, Bernasconi N (2009) Longitudinal and cross-sectional analysis of atrophy in pharmacoresistant temporal lobe epilepsy. *Neurology* 72(20):1747–1754. <https://doi.org/10.1212/01.wnl.0000345969.57574.f5>
 22. Bernhardt BC, Kim H, Bernasconi N (2013) Patterns of subregional mesiotemporal disease progression in temporal lobe epilepsy. *Neurology* 81(21):1840–1847
 23. Bernhardt BC, Hong SJ, Bernasconi A, Bernasconi N (2015) Magnetic resonance imaging pattern learning in temporal lobe epilepsy: classification and prognostics. *Annal Neurol*

- 77(3):436–446. <https://doi.org/10.1002/ana.24341>
24. Kim H, Mansi T, Bernasconi N (2013) Disentangling hippocampal shape anomalies in epilepsy. *Front Neurol* 4. <https://doi.org/10.3389/fneur.2013.00131>
25. Bernhardt BC, Bernasconi A, Liu M, Hong SJ, Caldairou B, Goubran M, Guiot MC, Hall J, Bernasconi N (2016) The spectrum of structural and functional imaging abnormalities in temporal lobe epilepsy. *Ann Neurol* 80(1):142–153. <https://doi.org/10.1002/ana.24691>
26. Yang J, Duncan JS (2004) 3D image segmentation of deformable objects with joint shape-intensity prior models using level sets. *Med Image Anal* 8(3):285–294. <https://doi.org/10.1016/j.media.2004.06.008>
27. Pitiot A, Delingette H, Thompson PM, Ayache N (2004) Expert knowledge-guided segmentation system for brain MRI. *Neuroimage* 23:S85–S96. <https://doi.org/10.1016/j.neuroimage.2004.07.040>
28. Duchesne S, Pruessner JC, Collins DL (2002) Appearance-based segmentation of medial temporal lobe structures. *Neuroimage* 17(2):515–531. <https://doi.org/10.1006/nimg.2002.1188>
29. Khan AR, Wang L, Beg MF (2008) FreeSurfer-initiated fully-automated subcortical brain segmentation in MRI using large deformation diffeomorphic metric mapping. *Neuroimage* 41(3):735–746. <https://doi.org/10.1016/j.neuroimage.2008.03.024>
30. Wang H, Das SR, Suh JW, Altinay M, Pluta J, Craige C, Avants B, Yushkevich PA (2011) A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation. *Neuroimage* 55(3):968–985. <https://doi.org/10.1016/j.neuroimage.2011.01.006>
31. Aljabar P, Heckemann RA, Hammers A, Hajnal JV, Rueckert D (2009) Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *Neuroimage* 46(3):726–738. <https://doi.org/10.1016/j.neuroimage.2009.02.018>
32. Collins DL, Pruessner JC (2010) Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. *Neuroimage* 52(4):1355–1366. <https://doi.org/10.1016/j.neuroimage.2010.04.193>
33. Kulaga-Yoskovitz J, Bernhardt BC, Hong SJ, Mansi T, Liang KE, van der Kouwe AJW, Smallwood J, Bernasconi A, Bernasconi N (2015) Multi-contrast submillimetric 3 Tesla hippocampal subfield segmentation protocol and dataset. *Sci Data* 2:150059–150059. <https://doi.org/10.1038/sdata.2015.59>
34. Iglesias JE, Augustinack JC, Nguyen K, Player CM, Player A, Wright M, Roy N, Frosch MP, McKee AC, Wald LL (2015) A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: application to adaptive segmentation of in vivo MRI. *Neuroimage* 115:117–137
35. Pipitone J, Park MTM, Winterburn J, Lett TA, Lerch JP, Pruessner JC, Lepage M, Voineskos AN, Chakravarty MM, Initiative ADN (2014) Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. *Neuroimage* 101:494–512
36. Van Leemput K, Bakkour A, Benner T, Wiggins G, Wald LL, Augustinack J, Dickerson BC, Golland P, Fischl B (2009) Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. *Hippocampus* 19(6):549–557
37. Kim H, Mansi T, Bernasconi N, Bernasconi A (2011) Robust surface-based multi-template automated algorithm to segment healthy and pathological hippocampi. In: Medical image computing and computer-assisted intervention—MICCAI 2011. Lecture notes in computer science, vol 6893, pp 445–453. https://doi.org/10.1007/978-3-642-2362-6-6_55
38. Caldairou B, Bernhardt BC, Kulaga-Yoskovitz J, Kim H, Bernasconi N, Bernasconi A (2016) A surface patch-based segmentation method for hippocampal subfields. In: International conference on medical image computing and computer-assisted intervention—MICCAI 2016. Lecture notes in computer science, vol 9901, pp 379–387. https://doi.org/10.1007/978-3-319-46723-8_44
39. Keihaninejad S, Heckemann RA, Gousias IS, Hajnal JV, Duncan JS, Aljabar P, Rueckert D, Hammers A (2012) Classification and lateralization of temporal lobe epilepsies with and without hippocampal atrophy based on whole-brain automatic MRI segmentation. *PLoS One* 7(4):e33096
40. Hadar PN, Kini LG, Coto C, Piskin V, Callans LE, Chen SH, Stein JM, Das SR, Yushkevich PA, Davis KA (2018) Clinical validation of

- automated hippocampal segmentation in temporal lobe epilepsy. *NeuroImage Clin* 20: 1139–1147
41. Mahmoudi F, Elisevich K, Bagher-Ebadian H, Nazem-Zadeh MR, Davoodi-Bojd E, Schwalb JM, Kaur M, Soltanian-Zadeh H (2018) Data mining MR image features of select structures for lateralization of mesial temporal lobe epilepsy. *PLoS One* 13(8): e0199137
 42. Beheshti I, Sone D, Maikusa N, Kimura Y, Shigemoto Y, Sato N, Matsuda H (2020) FLAIR-wise machine-learning classification and lateralization of MRI-negative 18F-FDG PET-positive temporal lobe epilepsy. *Front Neurol* 11(1433). <https://doi.org/10.3389/fneur.2020.580713>
 43. Beheshti I, Sone D, Maikusa N, Kimura Y, Shigemoto Y, Sato N, Matsuda H (2021) Accurate lateralization and classification of MRI-negative 18F-FDG-PET-positive temporal lobe epilepsy using double inversion recovery and machine-learning. *Comput Biol Med* 137:104805. <https://doi.org/10.1016/j.combiomed.2021.104805>
 44. Caldairou B, Foit NA, Mutti C, Fadaie F, Gill R, Lee HM, Demerath T, Urbach H, Schulze-Bonhage A, Bernasconi A (2021) An MRI-based machine learning prediction framework to lateralize hippocampal sclerosis in patients with temporal lobe epilepsy. *Neurology* 97(16):e1583–e1593
 45. Manjon JV, Romero JE, Coupe P (2020) DeepHIPS: a novel deep learning based hippocampus subfield segmentation method. arXiv preprint arXiv:200111789
 46. Zhu H, Shi F, Wang L, Hung SC, Chen MH, Wang S, Lin W, Shen D (2019) Dilated dense U-net for infant hippocampus subfield segmentation. *Front Neuroinform* 13(30). <https://doi.org/10.3389/fninf.2019.00030>
 47. Goubran M, Ntiri EE, Akhavein H, Holmes M, Nestor S, Ramirez J, Adamo S, Ozzoude M, Scott C, Gao F, Martel A, Swardfager W, Masellis M, Swartz R, MacIntosh B, Black SE (2020) Hippocampal segmentation for brains with extensive atrophy using three-dimensional convolutional neural networks. *Hum Brain Mapp* 41(2): 291–308. <https://doi.org/10.1002/hbm.24811>
 48. Gleichgerrcht E, Munsell BC, Alhusaini S, Alvim MK, Bargalló N, Bender B, Bernasconi A, Bernasconi N, Bernhardt B, Blackmon K (2021) Artificial intelligence for classification of temporal lobe epilepsy with ROI-level MRI data: a worldwide ENIGMA-Epilepsy study. *NeuroImage Clin* 31:102765
 49. Cohen-Gadol AA, Özduyan K, Bronen RA, Kim JH, Spencer DD (2004) Long-term outcome after epilepsy surgery for focal cortical dysplasia. *J Neurosurg* 101(1):55–65. <https://doi.org/10.3171/jns.2004.101.1.0055>
 50. Krsek P, Maton B, Korman B, Pacheco-Jacome E, Jayakar P, Dunoyer C, Rey G, Morrison G, Ragheb J, Vinters HV, Resnick T, Duchowny M (2008) Different features of histopathological subtypes of pediatric focal cortical dysplasia. *Ann Neurol* 63(6):758–769. <https://doi.org/10.1002/ana.21398>
 51. Krsek P, Maton B, Jayakar P, Dean P, Korman B, Rey G, Dunoyer C, Pacheco-Jacome E, Morrison G, Ragheb J, Vinters HV, Resnick T, Duchowny M (2009) Incomplete resection of focal cortical dysplasia is the main predictor of poor postsurgical outcome. *Neurology* 72(3):217–223. <https://doi.org/10.1212/01.wnl.0000334365.22854.d3>
 52. Widdess-Walsh P, Jeha L, Nair D, Kotagal P, Bingaman W, Najm I (2007) Subdural electrode analysis in focal cortical dysplasia: predictors of surgical outcome. *Neurology* 69(7):660–667
 53. Hedegård E, Bjellvi J, Edelvik A, Rydenhag B, Flink R, Malmgren K (2014) Complications to invasive epilepsy surgery workup with subdural and depth electrodes: a prospective population-based observational study. *J Neurol Neurosurg Psychiatry* 85(7):716–720
 54. Fauser S, Schulze-Bonhage A, Honegger J, Carmona H, Huppertz HJ, Pantazis G, Rona S, Bast T, Strobl K, Steinhoff BJ, Korinthenberg R, Rating D, Volk B, Zentner J (2004) Focal cortical dysplasias: surgical outcome in 67 patients in relation to histological subtypes and dual pathology. *Brain* 127(11):2406–2418. <https://doi.org/10.1093/brain/awh277>
 55. Cascino GD (2004) Surgical treatment for epilepsy. *Epilepsy Res* 60(2):179–186. <https://doi.org/10.1016/j.epilepsyres.2004.07.003>
 56. Focke NK, Symms MR, Burdett JL, Duncan JS (2008) Voxel-based analysis of whole brain FLAIR at 3T detects focal cortical dysplasia. *Epilepsia* 49(5):786–793
 57. Rugg-Gunn F, Eriksson S, Boulby P, Symms M, Barker G, Duncan J (2003)

- Magnetization transfer imaging in focal epilepsy. *Neurology* 60(10):1638–1645
58. Rugg-Gunn F, Boulby P, Symms M, Barker G, Duncan J (2005) Whole-brain T2 mapping demonstrates occult abnormalities in focal epilepsy. *Neurology* 64(2):318–325
 59. Salmenpera TM, Symms MR, Rugg-Gunn FJ, Boulby PA, Free SL, Barker GJ, Yousry TA, Duncan JS (2007) Evaluation of quantitative magnetic resonance imaging contrasts in MRI-negative refractory focal epilepsy. *Epilepsia* 48(2):229–237
 60. Bernasconi A, Antel SB, Collins DL, Bernasconi N, Olivier A, Dubeau F, Pike GB, Andermann F, Arnold DL (2001) Texture analysis and morphological processing of magnetic resonance imaging assist detection of focal cortical dysplasia in extra-temporal partial epilepsy. *Ann Neurol* 49(6):770–775. <https://doi.org/10.1002/ana.1013>
 61. Colliot O, Antel SB, Naessens VB, Bernasconi N, Bernasconi A (2006) In vivo profiling of focal cortical dysplasia on high-resolution MRI with computational models. *Epilepsia* 47(1):134–142. <https://doi.org/10.1111/j.1528-1167.2006.00379.x>
 62. Huppertz HJ, Grimm C, Fauser S, Kassubek J, Mader I, Hochmuth A, Spreer J, Schulze-Bonhage A (2005) Enhanced visualization of blurred gray-white matter junctions in focal cortical dysplasia by voxel-based 3D MRI analysis. *Epilepsy Res* 67(1–2):35–50
 63. Antel SB, Li LM, Cendes F, Collins DL, Kearney RE, Shinghal R, Arnold DL (2002) Predicting surgical outcome in temporal lobe epilepsy patients using MRI and MRSI. *Neurology* 58(10):1505–1512. <https://doi.org/10.1212/wnl.58.10.1505>
 64. Antel SB, Collins DL, Bernasconi N, Andermann F, Shinghal R, Kearney RE, Arnold DL, Bernasconi A (2003) Automated detection of focal cortical dysplasia lesions using computational models of their MRI characteristics and texture analysis. *Neuroimage* 19(4):1748–1759. [https://doi.org/10.1016/S1053-8119\(03\)00226-X](https://doi.org/10.1016/S1053-8119(03)00226-X)
 65. Adler S, Wagstyl K, Gunny R, Ronan L, Carmichael D, Cross JH, Fletcher PC, Baldeweg T (2016) Novel surface features for automated detection of focal cortical dysplasias in paediatric epilepsy. *Neuroimage Clin* 14:18–27. <https://doi.org/10.1016/j.nicl.2016.12.030>
 66. Gill RS, Hong SJ, Fadaie F, Caldairou B, Bernhardt B, Bernasconi N, Bernasconi A (2017) Automated detection of epileptogenic cortical malformations using multimodal MRI. In: Deep learning in medical image analysis and multimodal learning for clinical decision support: third international workshop, DLMIA 2017, ML-CDS 2017. Lecture notes in computer science, vol 10553, pp 349–356. https://doi.org/10.1007/978-3-319-67558-9_40
 67. Hong SJ, Kim H, Schrader D, Bernasconi N, Bernhardt B, Bernasconi A (2014) Automated detection of cortical dysplasia type II in MRI-negative epilepsy. *Neurology* 83(1):48–55. <https://doi.org/10.1212/WNL.0000000000000543>
 68. Jin B, Krishnan B, Adler S, Wagstyl K, Hu W, Jones S, Najm I, Alexopoulos A, Zhang K, Zhang J (2018) Automated detection of focal cortical dysplasia type II with surface-based magnetic resonance imaging postprocessing and machine learning. *Epilepsia* 59(5):982–992
 69. Tan YL, Kim H, Lee S, Tihan T, Ver Hoef L, Mueller SG, Barkovich AJ, Xu D, Knowlton R (2018) Quantitative surface analysis of combined MRI and PET enhances detection of focal cortical dysplasias. *Neuroimage* 166:10–18
 70. Snyder K, Whitehead EP, Theodore WH, Zaghloul KA, Inati SJ, Inati SK (2021) Distinguishing type II focal cortical dysplasias from normal cortex: a novel normative modeling approach. *Neuroimage Clin* 30:102565
 71. Kini LG, Gee JC, Litt B (2016) Computational analysis in epilepsy neuroimaging: a survey of features and methods. *Neuroimage Clin* 11:515–529
 72. Spitzer H, Ripart M, Whitaker K, Napolitano A, De Palma L, De Benedictis A, Foldes S, Humphreys Z, Zhang K, Hu W, Mo J, Likeman M, Davies S, Guttler C, Lenge M, Cohen NT, Tang Y, Wang S, Chari A, Tisdall M, Bargallo N, Conde-Blanco E, Pariente JC, Pascual-Diaz S, Delgado-Martínez I, Pérez-Enríquez C, Lagorio I, Abela E, Mullatti N, O’Muircheartaigh J, Vecchiato K, Liu Y, Caligiuri M, Sinclair B, Vivash L, Willard A, Kandasamy J, McLellan A, Sokol D, Semmelroch M, Kloster A, Opheim G, Ribeiro L, Yasuda C, Rossi-Espagnet C, Zhang K, Hamandi K, Tietze A, Barba C, Guerrini R, Gaillard WD, You X, Wang I, González-Ortiz S, Severino M, Striano P, Tortora D, Kalviainen R, Gambardella A, Labate A, Desmond P, Lui E, O’Brien T, Shetty J, Jackson G, Duncan J, Winston G, Pinborg L, Cendes F, Theis FJ, Shinohara RT, Cross JH, Baldeweg T, Adler S, Wagstyl K (2021) Interpretable surface-based detection of focal cortical dysplasias: a MELD study.

- m e d R x i v . [h t t p s : / / d o i . org/10.1101/2021.12.13.21267721](https://doi.org/10.1101/2021.12.13.21267721)
73. Najm IM, Sarnat HB, Blümcke I (2018) Review: the international consensus classification of Focal Cortical Dysplasia—a critical update 2018. *Neuropathol Appl Neurobiol* 44(1):18–31. <https://doi.org/10.1111/nan.12462>
 74. Iflland PH, Crino PB (2017) Focal cortical dysplasia: gene mutations, cell signaling, and therapeutic implications. *Annu Rev Pathol* 12(1):547–571. <https://doi.org/10.1146/annurev-pathol-052016-100138>
 75. Marsan E, Baulac S (2018) Review: mechanistic target of rapamycin (mTOR) pathway, focal cortical dysplasia and epilepsy. *Neuropathol Appl Neurobiol* 44(1):6–17. <https://doi.org/10.1111/nan.12463>
 76. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. *Med Image Anal* 42:60–88
 77. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 25(1):44–56
 78. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge, MA
 79. Dev KB, Jogi PS, Niyas S, Vinayagamani S, Kesavadas C, Rajan J (2019) Automatic detection and localization of Focal Cortical Dysplasia lesions in MRI using fully convolutional neural network. *Biomed Signal Process Control* 52:218–225
 80. Thomas E, Pawan S, Kumar S, Horo A, Niyas S, Vinayagamani S, Kesavadas C, Rajan J (2020) Multi-res-attention UNet: a CNN model for the segmentation of focal cortical dysplasia lesions from magnetic resonance images. *IEEE J Biomed Health Inf* 25(5):1724–1734
 81. Wang H, Ahmed SN, Mandal M (2020) Automated detection of focal cortical dysplasia using a deep convolutional neural network. *Comput Med Imaging Graph* 79:101662
 82. Gill RS, Lee HM, Caldaírou B, Hong SJ, Barba C, Deleo F, D'Incerti L, Coelho VCM, Lenge M, Semmelroch M (2021) Multicenter validation of a deep learning detection algorithm for focal cortical dysplasia. *Neurology* 97(16):e1571–e1582
 83. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S (2017) Leveraging uncertainty information from deep neural networks for disease detection. *Sci Rep* 7(1):17816. <https://doi.org/10.1038/s41598-017-17876-z>
 84. Gal Y, Ghahramani Z (2015) Bayesian convolutional neural networks with Bernoulli approximate variational inference. arXiv preprint arXiv:150602158
 85. Smolyansky ED, Hakeem H, Ge Z, Chen Z, Kwan P (2021) Machine learning models for decision support in epilepsy management: a critical review. *Epilepsy Behav* 123:108273
 86. Petrovski S, Szoekie CE, Sheffield LJ, D'souza W, Huggins RM, O'brien TJ (2009) Multi-SNP pharmacogenomic classifier is superior to single-SNP models for predicting drug outcome in complex diseases. *Pharmacogenet Genomics* 19(2):147–152
 87. Shazadi K, Petrovski S, Roten A, Miller H, Huggins RM, Brodie MJ, Pirmohamed M, Johnson MR, Marson AG, O'Brien TJ (2014) Validation of a multigenic model to predict seizure control in newly treated epilepsy. *Epilepsy Res* 108(10):1797–1805
 88. Silva-Alves MS, Secolin R, Carvalho BS, Yasuda CL, Bilevicius E, Alvim MK, Santos RO, Maurer-Morelli CV, Cendes F, Lopes-Cendes I (2017) A prediction algorithm for drug response in patients with mesial temporal lobe epilepsy based on clinical and genetic information. *PLoS One* 12(1):e0169214
 89. An S, Malhotra K, Dilley C, Han-Burgess E, Valdez JN, Robertson J, Clark C, Westover MB, Sun J (2018) Predicting drug-resistant epilepsy—a machine learning approach based on administrative claims data. *Epilepsy Behav* 89:118–125
 90. Devinsky O, Dilley C, Ozery-Flato M, Aharonov R, Goldschmidt Y, Rosen-Zvi M, Clark C, Fritz P (2016) Changing the approach to treatment choice in epilepsy using big data. *Epilepsy Behav* 56:32–37
 91. Delen D, Davazdahemami B, Eryarsoy E, Tomak L, Valluru A (2020) Using predictive analytics to identify drug-resistant epilepsy patients. *Health Inf J* 26(1):449–460
 92. Yao L, Cai M, Chen Y, Shen C, Shi L, Guo Y (2019) Prediction of antiepileptic drug treatment outcomes of patients with newly diagnosed epilepsy by machine learning. *Epilepsy Behav* 96:92–97
 93. Memarian N, Kim S, Dewar S, Engel Jr J, Staba RJ (2015) Multimodal data and machine learning for surgery outcome prediction in complicated cases of mesial temporal lobe epilepsy. *Comput Biol Med* 64:67–78
 94. Armañanzas R, Alonso-Nanclares L, DeFelipe-Oroquieta J, Kastanauskaitė A, de Sola RG, DeFelipe J, Bielza C, Larrañaga P (2013) Machine learning approach for the

- outcome prediction of temporal lobe epilepsy surgery. *PLoS One* 8(4):e62819
95. Bernhardt B, Hong SJ, Bernasconi A, Bernasconi N (2013) Imaging structural and functional brain networks in temporal lobe epilepsy. *Front Hum Neurosci* 7(624). <https://doi.org/10.3389/fnhum.2013.00624>
96. Caciagli L, Bernhardt BC, Hong SJ, Bernasconi A, Bernasconi N (2014) Functional network alterations and their structural substrate in drug-resistant epilepsy. *Front Neurosci* 8:411
97. Munsell BC, Wee CY, Keller SS, Weber B, Elger C, da Silva LAT, Nesland T, Styner M, Shen D, Bonilha L (2015) Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. *Neuroimage* 118:219–230
98. Taylor PN, Sinha N, Wang Y, Vos SB, De Tisi J, Misericocchi A, McEvoy AW, Winston GP, Duncan JS (2018) The impact of epilepsy surgery on the structural connectome and its relation to outcome. *Neuroimage Clin* 18: 202–214
99. He X, Doucet GE, Pustina D, Sperling MR, Sharan AD, Tracy JI (2017) Presurgical thalamic hubness predicts surgical outcome in temporal lobe epilepsy. *Neurology* 88(24): 2285–2293
100. Larivière S, Weng Y, Vos de Wael R, Royer J, Frauscher B, Wang Z, Bernasconi A, Bernasconi N, Schrader DV, Zhang Z, Bernhardt BC (2020) Functional connectome contractions in temporal lobe epilepsy: microstructural underpinnings and predictors of surgical outcome. *Epilepsia* 61(6): 1221–1233. <https://doi.org/10.1111/epi.16540>
101. Gleichgerrcht E, Keller SS, Drane DL, Munsell BC, Davis KA, Kaestner E, Weber B, Krantz S, Vandergift WA, Edwards JC (2020) Temporal lobe epilepsy surgical outcomes can be inferred based on structural connectome hubs: a machine learning study. *Ann Neurol* 88(5):970–983
102. Sinha N, Wang Y, da Silva NM, Misericocchi A, McEvoy AW, de Tisi J, Vos SB, Winston GP, Duncan JS, Taylor PN (2021) Structural brain network abnormalities and the probability of seizure recurrence after epilepsy surgery. *Neurology* 96(5):e758–e771
103. Lo A, Chernoff H, Zheng T, Lo SH (2015) Why significant variables aren't automatically good predictors. *Proc Natl Acad Sci USA* 112(45):13892–13897. <https://doi.org/10.1073/pnas.1518285112>
104. Arbabshirani MR, Plis S, Sui J, Calhoun VD (2017) Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *Neuroimage* 145:137–165. <https://doi.org/10.1016/j.neuroimage.2016.02.079>
105. Colombo N, Tassi L, Deleo F, Citterio A, Brammerio M, Mai R, Sartori I, Cardinale F, Lo Russo G, Spreafico R (2012) Focal cortical dysplasia type IIa and IIb: MRI aspects in 118 cases proven by histopathology. *Neuroradiology* 54(10):1065–1077. <https://doi.org/10.1007/s00234-012-1049-1>
106. Gross RE, Stern MA, Willie JT, Fasano RE, Saindane AM, Soares BP, Pedersen NP, Drane DL (2018) Stereotactic laser amygdalohippocampotomy for mesial temporal lobe epilepsy. *Ann Neurol* 83(3):575–587. <https://doi.org/10.1002/ana.25180>
107. Hong SJ, Lee HM, Gill R, Crane J, Sziklas V, Bernhardt BC, Bernasconi N, Bernasconi A (2019) A connectome-based mechanistic model of focal cortical dysplasia. *Brain* 142(3):688–699. <https://doi.org/10.1093/brain/awz009>
108. Lee HM, Gill RS, Fadaie F, Cho KH, Guiot MC, Hong SJ, Bernasconi N, Bernasconi A (2020) Unsupervised machine learning reveals lesional variability in focal cortical dysplasia at mesoscopic scale. *Neuroimage Clin* 28:102438. <https://doi.org/10.1016/j.nicl.2020.102438>
109. Margerison J, Corsellis J (1966) Epilepsy and the temporal lobes: a clinical, electroencephalographic and neuropathologic study of the brain in epilepsy, with particular reference to the temporal lobes. *Brain* 89(3):499–530. <https://doi.org/10.1093/brain/89.3.499>
110. De Lanerolle NC, Kim JH, Williamson A, Spencer SS, Zaveri HP, Eid T, Spencer DD (2003) A retrospective analysis of hippocampal pathology in human temporal lobe epilepsy: evidence for distinctive patient subcategories. *Epilepsia* 44(5):677–687
111. Blümcke I, Pauli E, Clusmann H, Schramm J, Becker A, Elger C, Merschhemke M, Meencke HJ, Lehmann T, von Deimling A (2007) A new clinico-pathological classification system for mesial temporal sclerosis. *Acta Neuropathol* 113(3):235–244
112. Reyes A, Kaestner E, Bahrami N, Balachandra A, Hegde M, Paul BM, Hermann B, McDonald CR (2019) Cognitive phenotypes in temporal lobe epilepsy are associated with distinct patterns of white matter network abnormalities. *Neurology* 92(17):e1957–e1968. <https://doi.org/10.1212/wnl.0000000000007370>

113. Rodríguez-Cruces R, Bernhardt BC, Concha L (2020) Multidimensional associations between cognition and connectome organization in temporal lobe epilepsy. *Neuroimage* 213:116706. <https://doi.org/10.1016/j.neuroimage.2020.116706>
114. Lee HM, Fadaie F, Gill R, Caldairou B, Sziklas V, Crane J, Hong SJ, Bernhardt BC, Bernasconi A, Bernasconi N (2021) Decomposing MRI phenotypic heterogeneity in epilepsy: a step towards personalized classification. *Brain* 145(3):897–908. <https://doi.org/10.1093/brain/awab425>
115. Arnatkevičiūtė A, Fulcher BD, Fornito A (2019) A practical guide to linking brain-wide gene expression and neuroimaging data. *Neuroimage* 189:353–367. <https://doi.org/10.1016/j.neuroimage.2019.01.011>
116. Markello RD, Arnatkevičiūtė A, Poline JB, Fulcher BD, Fornito A, Misic B (2021) Standardizing workflows in imaging transcriptomics with the abagen toolbox. *Elife* 10: e72129
117. Lipton ZC (2018) The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* 16(3):31–57

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made. The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 28

Machine Learning in Multiple Sclerosis

Bas Jasperse and Frederik Barkhof

Abstract

Multiple sclerosis (MS) is characterized by inflammatory activity and neurodegeneration, leading to the accumulation of damage to the central nervous system resulting in the accumulation of disability. MRI depicts an important part of the pathology of this disease and therefore plays a key part in diagnosis and disease monitoring. Still, major challenges exist with regard to the differential diagnosis, adequate monitoring of disease progression, quantification of CNS damage, and prediction of disease progression. Machine learning techniques have been employed in an attempt to overcome these challenges. This chapter aims to give an overview of how machine learning techniques are employed in MS with applications for diagnostic classification, lesion segmentation, improved visualization of relevant brain pathology, characterization of neurodegeneration, and prognostic subtyping.

Key words Multiple sclerosis, Machine learning, Artificial intelligence, Deep learning, Neuroimaging

1 Introduction to MS

Multiple sclerosis (MS) is a neuroinflammatory disease of the central nervous system (CNS) affecting women more than men, usually starting in young adulthood with a prevalence of >100 per 100,000 individuals in the Western world and rising [1].

1.1 Disease Characteristics

The most striking feature is the appearance of focal inflammatory lesions in the CNS visible on MR imaging of the brain (Fig. 1) and/or spinal cord that may give rise to partially reversible loss of motor, sensory, and cognitive function depending on lesion location and the magnitude of damage to local nerve tissue. As the disease and resulting damage to the central nervous system accumulate, irreversible disability progresses over time. Although tempting, not all of the accumulated disability can be explained by focal inflammatory lesions [2]. Diffuse neurodegeneration in the CNS is another histopathological feature deemed responsible for gradually accumulating disability, especially in the later stages of the disease. This neurodegeneration is thought to result from a process

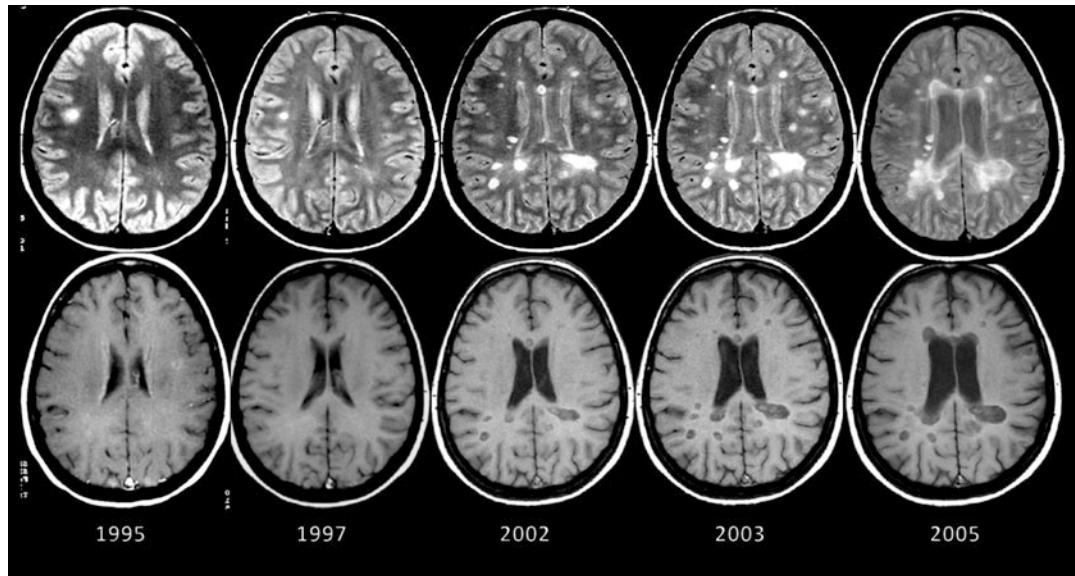


Fig. 1 PD-weighted MR images (top row) showing the occurrence of MS typical T2/PD hyperintense lesions over time. T1-weighted images (bottom row) showing enlargement of sulci and ventricles over time consistent with brain atrophy and hypointensity of multiple lesions due to local tissue loss. (Figure kindly provided by Dr. Alex Rovira (Hospital Universitari Vall d'Hebron, Barcelona))

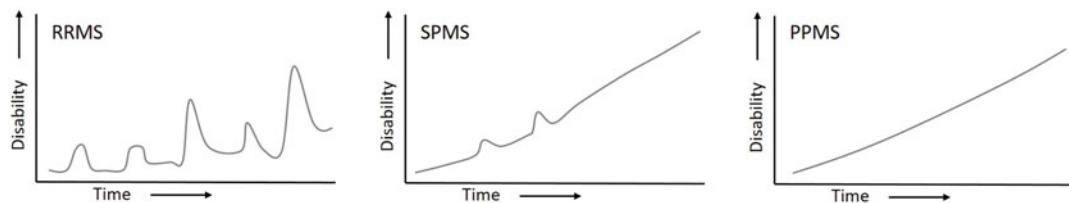


Fig. 2 MS subtypes based on the development of disability over time

that is partially separate from focal inflammation and can be visualized as initially subtle progressive brain atrophy on conventional MRI (Fig. 1) or by advanced MRI techniques that measure brain tissue integrity.

Based on the clinical course, MS is categorized in three main subtypes (Fig. 2). The most common subtype is relapsing remitting MS (RRMS), characterized by relapsing and remitting bouts of symptoms and limited disability. The RRMS subtype gradually transitions into the secondary progressive subtype (SPMS), characterized by gradually accumulating disability. Primary progressive MS (PPMS) is characterized by the gradual accumulation of disability from disease onset and is a common subtype in (male) patients with an older age at onset.

1.2 Treatment of MS

Suppression of CNS inflammation is the main target of treatment and has greatly improved over the past three decades. The earlier treatments are mainly based on molecules that suppress the CNS inflammatory response by interfering with cell signaling molecules that regulate the immune response (immunomodulation).

Interferon was the first immunomodulatory molecule to be approved for the treatment of RRMS in the last decade of the previous century, reducing relapse rate with approximately 30% as well as reducing the occurrence of inflammatory lesions on MRI [3, 4]. Other immunomodulatory molecules with similar efficacy have been developed and approved for the treatment of RRMS since the beginning of this century and include glatiramer acetate, dimethyl fumarate, and teriflunomide [5, 6]. These therapies are mostly well-tolerated with a low risk of serious adverse events.

Newer treatments are generally based on monoclonal antibodies that can directly block receptors on immune cells (immunosuppressive), disabling them to cause inflammation in the CNS, and include fingolimod, alemtuzumab, ocrelizumab, natalizumab, and siponimod [7–11]. These treatments are generally more effective than aforementioned immunomodulatory treatments with a reduction of the number of relapses with 50–80% and a more effective reduction of new active lesions on MRI. The downside of the latter treatments is the increased occurrence of more serious adverse events that include cardiovascular disease, autoimmune disease, and especially progressive multifocal leukoencephalopathy (PML). PML is caused by an infection of the CNS with the JC virus and the most dramatic and potentially lethal adverse event associated with the use of natalizumab, fingolimod, and, in very rare cases, dimethyl fumarate. Although comparatively less effective, the “immunomodulatory” treatments are recommended as “first-line” treatments due to their more favorable profile with regard to serious adverse events.

The quest for more effective and tolerable MS treatments that are also effective in patients with progressive MS is ongoing. New treatments that are currently being evaluated include vidofludimus calcium/IMU-838 [12], a dihydroorotate dehydrogenase inhibitor that attenuates pro-inflammatory cytokine release by B- and T-cells, and tolebrutinib, an inhibitor of the enzyme “Bruton’s tyrosine kinase” that drives CNS inflammation [13].

1.3 Diagnosis of MS

Proof of dissemination of inflammatory activity within the CNS in time and space is the underlying principle for diagnosing MS. This follows the successive bouts of focal inflammation in different parts of the CNS unique to the disease. Initially, these two criteria were fulfilled based on clinical course, in which at least two separate episodes of clinical disability (dissemination in time) related to separate locations in the CNS (dissemination in space) needed to be proven [14]. A first or multiple episodes of symptoms/signs

related to one location in the CNS is referred to as a clinically isolated syndrome (CIS). When a second episode occurs related to another location of the CNS, clinically definite MS (CDMS) can be diagnosed. Using this clinical diagnostic scheme, a definite diagnosis of MS could take years to be made and would take too long in the current era of effective treatment that need to be considered at an early stage of the disease. This has led to the incorporation of brain MRI findings in the diagnostic criteria following the same principles [15, 16]. In the diagnostic setting, the MR imaging protocol should at least include a FLAIR and T2 sequence of the brain to adequately detect and locate inflammatory lesions and a T1-weighted sequence of the brain after intravenous gadolinium contrast administration to detect active inflammatory lesions exhibiting leakage of contrast material into the local brain parenchyma. A T1 without contrast and DWI sequences of the brain is usually included for differential diagnostic purposes. T2-/PD-weighted and post-contrast T1-weighted sequences of the spinal cord are optional, when brain imaging is insufficient to make the diagnosis [17]. See Table 1 for an overview of the most frequently used MRI sequences for the diagnosis and monitoring of MS.

To fulfill the MRI criterion for dissemination in space, multiple inflammatory lesions should be demonstrated on brain or spinal cord MRI in two out of four typical CNS locations (i.e., juxta-/intra-cortical, periventricular, infratentorial, spinal cord). The dissemination in time criterion is fulfilled by demonstrating one or more new lesions on subsequent MRI scans and/or the simultaneous presence of lesions that do and do not enhance after gadolinium administration on any single scan. Further refinement of the diagnostic criteria has made it possible to make a diagnosis within 3–12 months of symptom onset for the vast majority cases with typical MS [18, 19].

1.4 Disease Monitoring

Disease progression is monitored by self-reporting of MS-associated symptoms, neurological assessment for MS-associated signs, and the detection of new lesions on MRI of the brain and/or spinal cord. Routine brain MRI is usually acquired each year and includes T2/PD and FLAIR sequences for the detection of new lesions. A DWI sequence of the brain is included to differentiate potential PML from MS lesions depending on the initiated treatment. More frequent MR imaging, T1-weighted post-contrast brain sequences, and imaging of the spinal cord are optional depending on clinical signs and symptoms and timing of treatment initiation [17]. See Table 1 for an overview of MRI sequences that are typically acquired for the monitoring of disease activity.

Table 1

Brief overview of sequences that are typically used in clinical practice for the diagnosis and monitoring of MS

	Diagnosis/purpose	Monitoring/purpose
Brain MRI		
Ax T1 (<3 mm 2D or 3D)	<i>Optional</i> Detection of T1 hypointense lesions	<i>Optional</i> Detection of T1 hypointense lesions
Ax T2 and PD (<3 mm)	<i>Recommended</i> Detection and localization of lesions (dissemination in space)	<i>Recommended</i> Detection of new lesions
FLAIR (preferably 3D with FS)	<i>Recommended</i> Detection and localization of lesions (dissemination in space)	<i>Recommended</i> Detection of new lesions
Ax T1 after contrast <3 mm 2D or 3D)	<i>Recommended</i> Detection of (in)active inflammation (dissemination in time)	<i>Optional</i> Detection of new active inflammation
DIR	<i>Optional</i> Improve detection of (juxta)cortical lesions	<i>Optional</i> Detection of new lesions
Ax DWI	<i>Optional</i> Characterization of lesions (differential diagnosis)	<i>Optional</i> Differentiation of MS versus PML lesions
Optic nerve MRI		
Ax/cor T2 FS or STIR (≤3 mm)	<i>Optional</i> Detection of optic neuritis	<i>Not required</i>
Ax/cor T1 after contrast (≤3 mm)	<i>Optional</i> Detection of active optic neuritis	<i>Not required</i>
Spinal cord MRI		
Sag T2 and PD (≤3 mm)	<i>Optional</i> Detection of spinal cord lesions (dissemination in space)	<i>Optional</i> Detection of new spinal cord lesions
Sag T1 after contrast (≤3 mm)	<i>Optional</i> Detection of active inflammation (dissemination in time)	<i>Optional</i> Detection of new active inflammation

For a detailed description see the 2021 MAGNIMS recommendations [17]. (Note: local preferences may vary)

Ax axial orientation, *Sag* sagittal orientation, *Cor* coronal orientation, *FS* fat suppression, *DWI* diffusion-weighted imaging, *PML* progressive multifocal leukoencephalopathy

1.5 Advanced MR Imaging Techniques

More advanced MRI techniques, like magnetic transfer ratio (MTR), diffusion tensor imaging (DTI), and resting state functional MRI (rsfMRI), are generally not used for the diagnosis and monitoring of MS patients in clinical practice as clinically relevant changes are hard to determine due to considerable biological and

technical (inter-/intra-scanner or inter-/intra-sequence) variability. These advanced sequences have been successfully used in controlled research settings to gain knowledge on the functional and structural dynamics of the MS disease process. MTR and DTI are mainly used to quantify microstructural integrity by measuring spin relaxation times and diffusion of protons within, in general, white matter tracts respectively. RsfMRI uses the BOLD effect to measure functional brain activity in the resting brain and/or in relation to specific tasks.

2 Machine Learning to Aid in the Differential Diagnosis of MS

Although the current diagnostic criteria are highly accurate and efficient in most cases of suspected MS, diagnostic challenges arise when atypical clinical and/or radiological findings occur that may represent other diseases that mimic multiple sclerosis. To aid in these diagnostic challenges, machine learning techniques have been employed in an attempt to distinguish MS from other diseases.

2.1 Differentiation of MS from Neuromyelitis Optica Spectrum Disorder

Neuromyelitis optica spectrum disorder (NMOSD) has previously been considered a variant of multiple sclerosis due to similarities in clinical presentation and presence of inflammatory lesions in the optic nerve, the spinal cord, and, especially in later stages, the brain. NMOSD has only recently been identified as a separate disease entity [20], especially with the identification of elevated antibodies against aquaporin-4, a water channel involved in water homeostasis in the CNS, and antibodies against myelin oligodendrocyte glycoprotein (MOG), a constituent of the normal myelin sheath. Although the clinical and radiological differences are known, the differential diagnosis remains a challenge due to the considerable overlap with MS.

Various machine learning models have been developed to differentiate between MS and NMOSD using decision trees based on expert findings of MRI of the orbits, brain, and spine [21], random forest analysis on radiomic features of brain lesions [22], CNN on brain MR images [23, 24], and LASSO binary logistic regression on the combination of radiomic features from spinal cord scans and clinical variables [25]. Performance of these models had AUCs varying between 0.712 and 0.935.

2.2 Differentiating MS from Other Diseases

A variety of other inflammatory autoimmune diseases and vascular diseases can present with similar brain MRI findings as MS. These diseases are usually easier to distinguish from MS using clinical variables such as age and disease course. However, MRI findings of the brain and spinal cord can still pose a challenge for radiologists who are not experienced with these pathologies.

Using support vector machine analysis on MRI-based radiomic features of brain lesions, Luo et al. created a model able to distinguish brain lesions in RRMS from systemic lupus erythematosus patient with an AUC of 0.967 [26].

In a broader effort, Rauschecker et al. [27] have created a machine learning model to provide a neuroradiological differential diagnosis for a range of brain diseases including MS. In their approach, they first detected and segmented brain lesions from brain MRI scans using a U-Net-based deep learning algorithm. They subsequently extracted 18 location-, spatial-, and signal-based quantitative imaging features using multiple pulse sequences from the segmented lesions. A Bayesian classifier was then used to combine these 18 image features with 5 clinical features for the prediction of the underlying brain disease. This classifier was able to make an accurate top three differential diagnosis in 91% of cases, with a similar performance as specialized academic neuroradiologists (86%, $P = 0.20$). More interestingly, this classifier outperformed neuroradiology fellows (77%, $P = 0.003$), general radiologists (57%, $P < 0.001$), and radiology residents (56%, $P < 0.001$). However, the datasets used were small (total $N = 86$ for training and $N = 92$ for testing, with N typically around 5 for each diagnostic class), and the performance for MS and related disorders like migraine was less favorable.

2.3 Future Considerations

Taken together, these studies show that machine learning has the capability to assist in the differential diagnosis of MS and can be especially helpful for radiologists that are not specialized in neuroradiology.

Most of the aforementioned ML models that could aid in differential diagnosis have focused on the differentiation between MS and NMOSD. Although interesting from a scientific point of view, this distinction is not the only diagnostic challenge from a clinical point of view. The main challenge for radiologists that are not experienced with these disease entities is the distinction between demyelinating lesions due to MS and vascular lesions and should be the focus of future studies.

Generalizability to the general population and MRI scanners is clearly the most important hurdle before these models can be introduced in clinical practice. In addition, these studies are generally limited to a small subset of differential diagnoses, which could lead to tunnel vision when relying on these tools in clinical practice.

3 Machine Learning for Lesion Segmentation and Quantification

Although lesions do not fully relate to the accumulation of clinical disability over time [2], lesion volume is still regarded as an important outcome measure in MS research and clinical trials, requiring accurate lesion segmentation. Manual lesion segmentation on MR images is highly labor-intensive and time-consuming, for which automated segmentation is an obvious solution, especially for 3D scans. Over the years, many (semi-)automated lesion segmentation techniques have been developed, including semi-automated seed growing and unsupervised K-means clustering techniques. In the recent years, convolutional neural networks have been shown to work particularly well in lesion segmentation tasks [28].

3.1 Cross-Sectional Lesion Segmentation

A large number of ML-based models have been developed that provide cross-sectional automated lesion segmentation in MS [29–38] using a variety of ML architecture designs. Critical evaluation and comparison of these large and increasing number of lesion segmentation methods is necessary to determine the best performing methods and their added value to existing methods using large test datasets made available in various challenges organized by the Medical Image Computing and Computer Assisted Intervention Society (MICCAI <http://www.miccai.org/>) and the International Symposium on Biomedical Imaging (ISBI, <https://biomedicalimaging.org/>) [28, 39]. Previous MS lesion segmentation challenges showed that segmentation algorithms could attain an average Dice score of 0.59 and an average surface distance of 0.91 for the segmentation of cross-sectional images in the MICCAI 2016 challenge [28] and an average Dice score of 0.670 and average symmetric surface distance of 2.16 for the segmentation of longitudinal MR images in the ISBI 2015 challenge [39]. Most of these algorithms required multiple input sequences, including T1, T2, PD, and/or FLAIR sequences, whereas only three algorithms required a single FLAIR sequence as input.

Besides segmentation performance, these methods need to be validated in real-world scenario with subjects scanned on MRI machines different from the original training dataset. To achieve this, some level of adjustment/optimization prior to implementation on a given dataset is generally needed. Weeda et al. [40] have compared methods with and without local optimization for *cross-sectional* segmentation of MS lesions using several freely available tools including LST [33], NicMSlesions [41], and BIANCA [31] (Fig. 3). Optimization to the local dataset improved performance for all these methods, while retraining with manually labelled representative MR images provided the best performance.

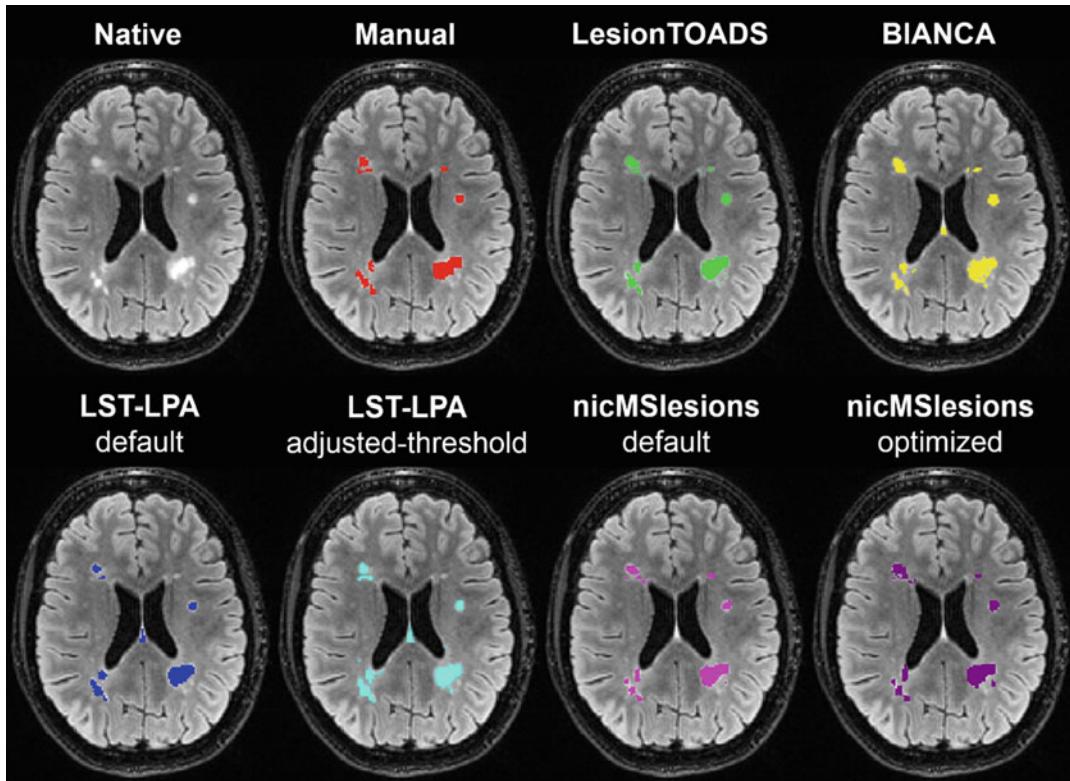


Fig. 3 Output examples of four lesion segmentation algorithms and manual segmentation overlaid over FLAIR images of the brain. (Figure adapted from Weeda et al. [40], reprinted with permission from Elsevier)

3.2 Detection of New MS Lesions

The detection of new lesions *longitudinally* is a highly important clinical monitoring task to demonstrate new inflammatory activity in the CNS that may prompt initiation or change of treatment for an individual MS patient. This requires tedious and time-consuming visual comparison of FLAIR images, especially in patients with a high number of confluent lesions. Initially, the aim of any treatment was to have no evidence of disease activity (NEDA). This proved to be unrealistic, as a low number of new lesions over time could be observed in patients treated with various treatment modalities [42]. Additional studies have shown that long-term clinical disability does not increase with two or less new lesions within 1 year and no contrast-enhancing lesions (minimal evidence of disease activity (MEDA)).

Various machine learning models have been created to detect *new* MS lesions on subsequent MR images based on fusion or subtraction of subsequent segmentation maps [33, 43–48] or end-to-end training of a combined registration and segmentation network on serial MRI scans [48]. Evaluation of these and new machine learning tools are expected following the recent MICCAI challenge (<https://portal.fli-iam.irisa.fr/msseg-2/>).

A recent study has compared the number of new lesions detected by visual assessment (highest sensitivity/accuracy: 69/67), automated assessment (highest sensitivity/accuracy: 84/64), and visual verification of the automated assessments (sensitivity/accuracy: 86/NA) on a single-center cohort of 100 MS patients [49]. The automated methods detected a higher number of new MS lesions than visual assessment. Visual verification of automated assessments revealed a high number of false positive new lesions when using automated assessments only and a high number of false negative new MS lesions with only visual assessments. Evidently, automated tools for new lesion detection require further development before they can be implemented in clinical practice *without supervision*. More importantly, this study showed that *visually supervised* automated methods are currently able to improve the detection of new MS lesions in current clinical practice. This would warrant clinical implementation, provided that the clinical tool allows swift and efficient visual supervision and correction and has a reasonable tradeoff between false negative and false positive rates erring slightly to the false positive side.

3.3 Clinical Implementation of ML Tools for Lesion Segmentation and Detection

Commercial image analysis packages meant for implementation in clinical care have incorporated automated lesion segmentation algorithms to provide cross-sectional and longitudinal assessments of lesion volume rather than (new) lesion counts. Although this may provide more precise monitoring of the patient's overall lesion burden, the utility of these tools should be critically evaluated on at least the following points: (1) knowledge of the robustness of the lesion segmentation algorithm to inter-scanner variability and various MR artifacts; (2) proven clinically relevant cut-off points of the provided measurements that are related to relevant future disability progression; and (3) implementation of a mandatory visual check of provided lesion counts and volumes.

4 Machine Learning to Improve Detection of Tissue Properties from Conventional MRI Sequences

MR imaging is the modality of choice for the diagnosis and monitoring of MS patients in clinical trials and daily clinical practice, due to its availability. Scan protocols in daily clinical practice are usually limited to the most essential conventional sequences to limit burden on patients and to limit financial cost. Machine learning can be employed to enhance these conventional sequences to visualize initially inconspicuous relevant tissue properties.

4.1 Synthetic DIR Sequences

Cortical lesions are an important part of MS pathology, specific to the disease, associated with disease progression [50, 51] and have recently been included in the radiological diagnostic criteria [52]. These cortical lesions are generally inconspicuous on commonly used FLAIR and T2-/PD-weighted MR images. The double inversion recovery MRI sequence (DIR) is uniquely capable of visualizing these cortical lesions by combined suppression of the MR signal from cerebrospinal fluid and white matter [53]. DIR sequences are generally not used in daily clinical practice or clinical trials due to the long acquisition time and lack of availability on most MR systems. Models based on generative adversarial networks have been trained to generate synthetic DIR images from conventional and routinely acquired T1, T2, and FLAIR images [54] and T1 and PD/T2 [55]. These synthetic DIR images were able to improve the detection of juxtacortical lesions (12.3 ± 10.8 vs 7.2 ± 5.6 , $P < 0.001$) [54] and cortical lesions ($N = 626$ vs 696) [56] compared to conventional MRI sequences. Although not as sensitive as the original DIR images, synthetic DIR images are sensitive enough to improve diagnosis and prognostication in routine clinical setting.

4.2 Prediction of Contrast-Enhancing Lesions

Besides a very low risk of nephrogenic systemic fibrosis [57], gadolinium-based contrast agents are generally safe when used for imaging purposes. However, gadolinium is known to accumulate in the brain after repeated IV gadolinium administrations. Although no adverse effects have been demonstrated to date, this is a cause for concern in the medical community as the long-term effects are still unknown. Because of this, prediction of the presence of active inflammatory contrast-enhancing lesions without the use of contrast agents is desirable. Using a large multicenter dataset, Narayana et al. have developed a deep learning model capable of predicting contrast-enhancing lesions using T1, T2, and FLAIR images with sensitivity and specificity of 78% and 73%, respectively, for patient-wise detection of enhancement using fivefold cross-validation [58].

4.3 Visualization of Tissue Myelin Content from MR Images

Demyelination is one of the pathological hallmarks of MS that cannot be directly quantified by MR imaging. In vivo quantification can be useful for monitoring inflicted damage by inflammation and the efficacy of myelin repair mechanisms. PET imaging is capable of visualizing and quantifying myelin using the radiotracer [(11)C] PIB [59], but is not generally available, expensive, and invasive. A recent study has used [(11)C]PIB PET images from MS patients to train a CF-SAGAN-based model to successfully predict myelin content changes from MTR, DTI, T2, and T1 MRI sequences [60] (Fig. 4).

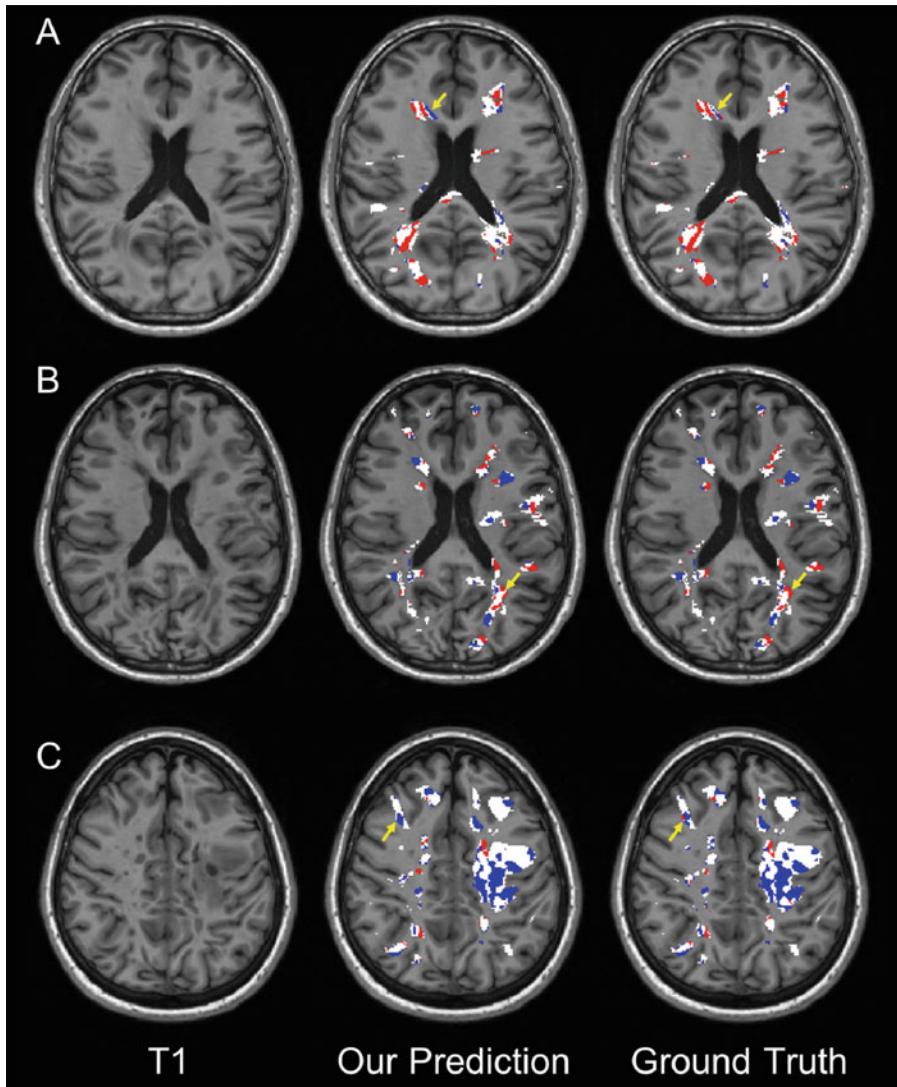


Fig. 4 Examples of lesional myelin content changes showing T1-weighted images (left column), the predicted change in myelin content by the MR-based model proposed by Wei et al. (middle column), and the ground truth change in myelin content based on [(11)C]PIB PET imaging (right column). Demyelinating (red) and remyelinating (in blue) voxels are indicated on top of the lesion mask (white). (Figure adapted from Wei et al. [60], reprinted with permission from Elsevier)

5 Machine Learning to Characterize Neurodegeneration in MS

In daily clinical practice, treatment changes in the course of the disease are mainly based on new inflammatory/demyelinating activity visible as new or enhancing lesions on brain MRI scans. In contrast, the partially unrelated but clinically relevant neurodegenerative aspect of the MS disease process is generally

underappreciated in monitoring and treatment decisions. The reason for this is the absence of simple, reliable, and easily interpretable measures that reflect the degree of neurodegeneration in individual patients.

Overall brain volume measured on MR images is currently the most important tool to quantify neurodegeneration in MS. However, brain volume measurements have not been implemented in routine clinical care as universal clinically relevant cut-off points for brain volume loss have not been identified due to considerable technical, biological, and, specifically, age-related variations [61].

5.1 Brain Age Determination from MR Images

Neurodegenerative processes are known to change the macroscopic structure of the brain with increasing age. Similar brain structure changes are observed as a result of various neurodegenerative brain diseases, including MS. Such MS-related atrophic changes occur at a faster pace as would be expected in normal aging individuals. This has given rise to the “brain age” paradigm, in which accelerated aging of the brain is considered as a marker of MS-related neurodegeneration [62]. Machine learning models based on large populations of healthy aging individuals have been developed to determine biological brain age from T1-weighted MR images of the brain [63–65]. Subtraction of this predicted brain age from the actual calendar age results in the brain-predicted age difference (brain-PAD) or brain age gap (BAG) as an indicator of premature aging of the brain. Key advantages of brain-PAD/BAG over brain volume measurement are that these measures incorporate image characteristics across the entire brain (not only the segmented brain tissue as in brain volume measurement), provide an intuitive easily interpretable metric, are more robust to acquisition-related image variations, and, most importantly, are specific for the individual patient by inherently adjusting for age. Initial studies on brain age in MS found that the estimated brain age is between 4 and 6 years higher than chronological age in comparison to healthy controls and that a higher relative brain age is associated with a higher degree of disability [65, 66]. A large retrospective multicenter study of brain age in MS showed that brain age is approximately 10 years higher than chronological age in MS, is increased in MS compared to HC, predicts current as well as future disability, and is mainly driven by brain atrophy [67] (Fig. 5). Recent developments in brain age models were made to reliably predict brain age using FLAIR sequences instead of the usual T1-weighted sequences (Colman et al., 2021; ISMRM 2022), ensuring flexible implementation in retrospective research settings and general clinical practice. Further studies are needed to further elucidate changes in brain age over time, the relationship of brain age with a wider range of measures of cognitive and physical disability, the influence of non-MS-related factors on brain age, the pathological substrate of brain

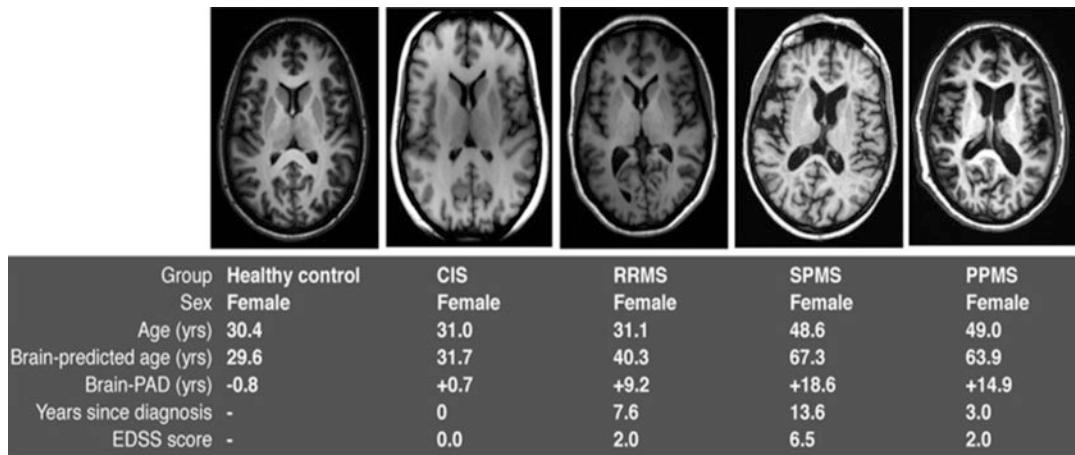


Fig. 5 Examples of increasing differences between brain-predicted age and chronological age (brain-PAD) in a healthy control, three RRMS onset patients with increasing disease durations, and a PPMS patient with a very high brain-PAD with relatively short time since diagnosis. (Figure adapted from Cole et al. [67], CC BY 4.0)

age in MS, and the effect of treatment on brain age. In the future, these brain age models may provide a useful clinical tool to quantify and monitor neurodegeneration in routine clinical care of MS patients.

5.2 Evolution of Brain Atrophy Over Time

Although overall progressive WM and GM atrophy is a well-known feature of MS, less is known on the evolution of atrophy in different brain regions over time. Event-based modelling [68, 69] has been used to elucidate the sequence in which GM atrophy affects various brain structures in repeated MRIs of 1417 subjects including healthy controls and all subtypes of MS [70, 71]. The posterior cingulate cortex and precuneus were the first regions to become atrophic, followed by the middle cingulate cortex, brainstem, and thalamus in patients with clinically isolated syndrome and relapse-onset MS. A similar pattern of sequential atrophy was found in PPMS with the involvement of the thalamus, cuneus, precuneus, and pallidum, followed by the brainstem and posterior cingulate cortex. Patients were then categorized according to the event stage defined by their individual atrophy pattern. Using a linear mixed effect model, progression of event stages was found to be related to the rate of disability progression proving that these atrophy stages represent clinically relevant GM pathology.

6 Machine Learning to Predict Disease Progression

The efficacy in reducing inflammatory activity, and thus preventing disability, varies across treatments and is generally speaking inversely related to side effects. Choosing the treatment with the

right tradeoff between efficacy and side effects is challenging as the disability accumulation over time can vary greatly among patients. Demographic variables, presence of oligoclonal bands, and the number of, especially infratentorial, T2 lesions at baseline brain MRI are known to be predictive of future disability progression and the likelihood of clinical relapse in the future [72]. Still, prediction of future disease progression remains a challenge in daily clinical practice especially when these risk factors are not unequivocally present. Several definitions of disease progression exist and include demonstration of short-term inflammatory activity (prediction of time to next relapse or progression from CIS to CDMS), changes in disability status using standardized clinical evaluations (EDSS progression or time to a certain clinical threshold), or progression from RRMS to SPMS.

6.1 Prediction of Disease Progression

ML techniques have successfully created models to predict worsening of disability based on CNN-based analysis of lesion maps, MR images, and age at baseline [73] and by combining clinical disability status and MRI-derived lesion volume and brain atrophy using SVM classifiers [74]. The latter study showed that the predictive properties of the SVM model improved when adding changes in MRI measurements over the first year.

A number of studies have successfully predicted a second relapse or conversion from CIS to CDMS by analyzing clinical and demographic data, lesion-specific quantitative geometric features, and gray matter-to-whole brain volume ratios using support vector machines [75]; clinical characteristics as well as global and local measures of GM/WM volume, lesion volume, and cortical thickness using support vector machines in combination with recursive feature elimination [76]; and lesion shape features derived from computer-assisted manual segmentation using a random forest classifier [77]. Pareto et al. created a model based on regional gray matter volume and T1 hypointensities obtained from the baseline T1-weighted MR images, but were not able to accurately predict conversion from CIS to CDMS [78].

6.2 Stratification of Patients at Risk of Disease Progression

Although the aforementioned models provide valuable insights into the predictive properties of clinical and radiological variables, the value to individual patients in daily clinical practice is still limited. An important downside of these models is the assumption that the predictive properties of baseline variables are monotonous among patients, whereas these predictive properties may well vary over time and between patients. Recent studies have applied the SuStaIn model [79] to identify MS subtypes based on clinical and radiological variables with the underlying assumption that these variables evolve over time. Using this technique on MRI-derived GM volumes in various brain regions, white matter volume, total brain lesion volume, and T1/T2 ratio within brain structures of

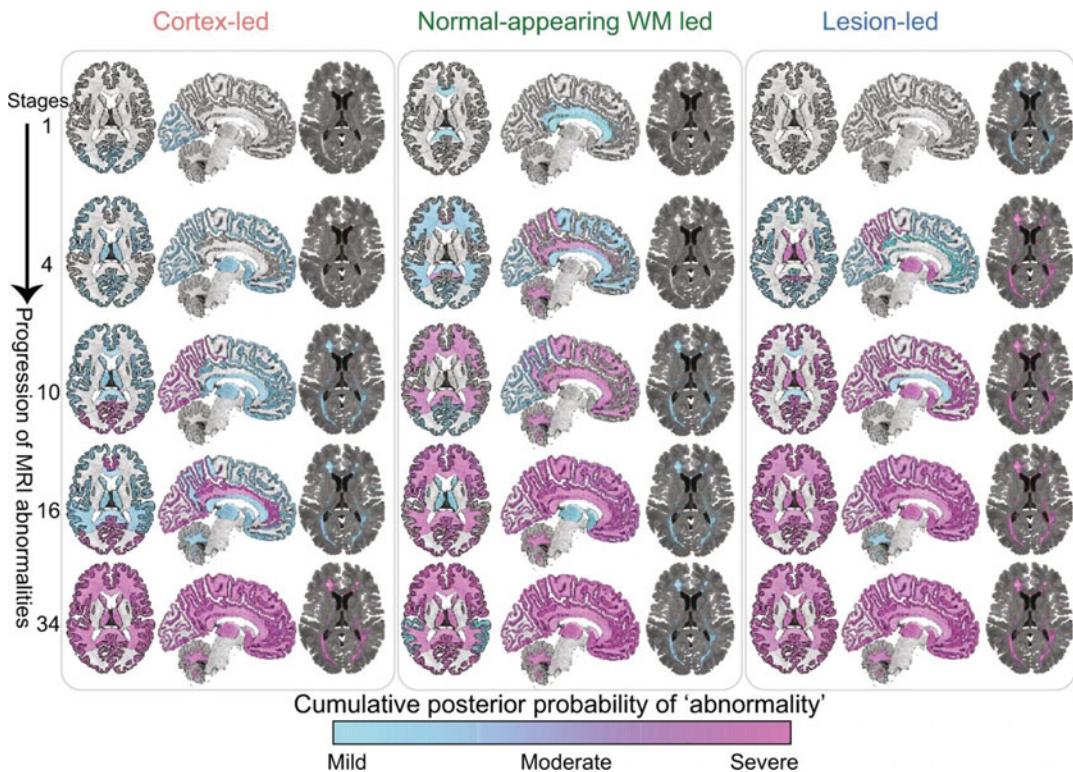


Fig. 6 Evolution of MRI abnormalities in each of the three MRI-based subtypes revealed by the SuStIn analysis by Eshaghi et al. For each subtype, the left two columns depict the probability of regional brain atrophy, and the right column depicts the probability of lesion occurrence in the various stages of MRI abnormality progression. (Figure adapted from Eshaghi et al. [81], CC BY 4.0)

6322 MS patients, Eshaghi et al. were able to define “cortex-led,” “normal-appearing white matter-led,” and “lesion-led MS” subtypes in the earliest stages of the disease [80, 81] (Fig. 6). Further analysis in the validation dataset ($N = 3068$) revealed that the lesion-led subtype had a significantly higher risk of disability progression, relapse rate, and treatment response in the following 24 weeks compared to the other two subtypes. Similar findings were made in a separate study on 425 MS patients analyzing GW matter volume in various brain regions, and T2 lesion volume using SuStIn revealed a subtype characterized by early deep GM atrophy and lesion appearance and a subtype characterized by early cortical GM volume loss that were consistent over time [82]. The subtype with early deep GM atrophy was associated with earlier disability progression and cognitive impairment compared to the subtype with earlier cortical volume loss. Taken together, these studies show that SuStIn modelling can reveal previously unknown subtypes of MS that are biologically and clinically relevant. The SuStIn models can be used to stratify individual patients and therefore has the potential for implementation in daily clinical practice after

adaptation of the model to include robust measurements that can be derived from MRI scans acquired in daily clinical practice.

7 Concluding Remarks

As covered in this chapter, ML techniques provide new insights and possibilities with regard to differential diagnosis, lesion segmentation and quantification, enhanced detection of relevant pathology on MRI, characterization of neurodegeneration, and prediction of disease progression in MS. In general, challenges still exist with regard to generalizability to the general population, robustness across images acquired from different MRI scanners, and validation that the ML technique provides biologically and clinically relevant information. Implementation of various ML tools in clinical practice is ongoing, but should provide insight in their robustness across scanners, clearly defined clinically relevant cut-off points for each provided outcome, and an efficient interface that allows the user to check the quality of the analyses when appropriate.

Acknowledgments

Frederik Barkhof is supported by the NIHR Biomedical Research Centre at UCLH.

Disclosures **Frederik Barkhof:** Steering committee or IDMC member for Biogen, Merck, Roche, Eisai, and Prothena. Consultant for Roche, Biogen, Merck, IXICO, Jansen, and Combinostics. Research agreements with Merck, Biogen, GE Healthcare, and Roche. Co-founder and shareholder of Queen Square Analytics Ltd.

References

- Walton C et al (2020) Rising prevalence of multiple sclerosis worldwide: insights from the Atlas of MS, third edition. *Mult Scler* 26: 1816–1821. <https://doi.org/10.1177/1352458520970841>
- Barkhof F (2002) The clinico-radiological paradox in multiple sclerosis revisited. *Curr Opin Neurol* 15:239–245. <https://doi.org/10.1097/00019052-200206000-00003>
- (1995) Interferon beta-1b in the treatment of multiple sclerosis: final outcome of the randomized controlled trial. The IFNB Multiple Sclerosis Study Group and The University of British Columbia MS/MRI Analysis Group. *Neurology* 45:1277–1285
- Jacobs LD et al (1996) Intramuscular interferon beta-1a for disease progression in relapsing multiple sclerosis. The Multiple Sclerosis Collaborative Research Group (MSCRG). *Ann Neurol* 39:285–294. <https://doi.org/10.1002/ana.410390304>
- Fox RJ et al (2012) Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. *N Engl J Med* 367:1087–1097. <https://doi.org/10.1056/NEJMoa1206328>
- O'Connor P et al (2011) Randomized trial of oral teriflunomide for relapsing multiple

- sclerosis. *N Engl J Med* 365:1293–1303. <https://doi.org/10.1056/NEJMoa1014656>
7. Kappos L et al (2010) A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. *N Engl J Med* 362:387–401. <https://doi.org/10.1056/NEJMoa0909494>
 8. Hauser SL et al (2017) Ocrelizumab versus interferon beta-1a in relapsing multiple sclerosis. *N Engl J Med* 376:221–234. <https://doi.org/10.1056/NEJMoa1601277>
 9. Polman CH et al (2006) A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. *N Engl J Med* 354:899–910. <https://doi.org/10.1056/NEJMoa044397>
 10. Cohen JA et al (2012) Alemtuzumab versus interferon beta 1a as first-line treatment for patients with relapsing-remitting multiple sclerosis: a randomised controlled phase 3 trial. *Lancet* 380:1819–1828. [https://doi.org/10.1016/S0140-6736\(12\)61769-3](https://doi.org/10.1016/S0140-6736(12)61769-3)
 11. Kappos L et al (2018) Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. *Lancet* 391:1263–1273. [https://doi.org/10.1016/S0140-6736\(18\)30475-6](https://doi.org/10.1016/S0140-6736(18)30475-6)
 12. Muehler A, Peelen E, Kohlhof H, Groppel M, Vitt D (2020) Vidofludimus calcium, a next generation DHODH inhibitor for the treatment of relapsing-remitting multiple sclerosis. *Mult Scler Relat Disord* 43:102129. <https://doi.org/10.1016/j.msard.2020.102129>
 13. Reich DS et al (2021) Safety and efficacy of tolebrutinib, an oral brain-penetrant BTK inhibitor, in relapsing multiple sclerosis: a phase 2b, randomised, double-blind, placebo-controlled trial. *Lancet Neurol* 20:729–738. [https://doi.org/10.1016/S1474-4422\(21\)00237-4](https://doi.org/10.1016/S1474-4422(21)00237-4)
 14. Poser CM et al (1983) New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Ann Neurol* 13:227–231. <https://doi.org/10.1002/ana.410130302>
 15. Barkhof F et al (1997) Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis. *Brain* 120(Pt 11):2059–2069. <https://doi.org/10.1093/brain/120.11.2059>
 16. McDonald WI et al (2001) Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Ann Neurol* 50: 121–127. <https://doi.org/10.1002/ana.1032>
 17. Wattjes MP et al (2021) 2021 MAGNIMS-CMSC-NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. *Lancet Neurol* 20:653–670. [https://doi.org/10.1016/S1474-4422\(21\)00095-8](https://doi.org/10.1016/S1474-4422(21)00095-8)
 18. Polman CH et al (2011) Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann Neurol* 69:292–302. <https://doi.org/10.1002/ana.22366>
 19. Swanton JK et al (2007) MRI criteria for multiple sclerosis in patients presenting with clinically isolated syndromes: a multicentre retrospective study. *Lancet Neurol* 6:677–686. [https://doi.org/10.1016/S1474-4422\(07\)70176-X](https://doi.org/10.1016/S1474-4422(07)70176-X)
 20. Wingerchuk DM, Lennon VA, Lucchinetti CF, Pittock SJ, Weinshenker BG (2007) The spectrum of neuromyelitis optica. *Lancet Neurol* 6: 805–815. [https://doi.org/10.1016/S1474-4422\(07\)70216-8](https://doi.org/10.1016/S1474-4422(07)70216-8)
 21. Clarke L et al (2021) MRI patterns distinguish AQP4 antibody positive neuromyelitis optica spectrum disorder from multiple sclerosis. *Front Neurol* 12:722237. <https://doi.org/10.3389/fneur.2021.722237>
 22. Huang J et al (2021) Multi-parametric MRI phenotype with trustworthy machine learning for differentiating CNS demyelinating diseases. *J Transl Med* 19:377. <https://doi.org/10.1186/s12967-021-03015-w>
 23. Hagiwara A et al (2021) Differentiation between multiple sclerosis and neuromyelitis optica spectrum disorders by multiparametric quantitative MRI using convolutional neural network. *J Clin Neurosci* 87:55–58. <https://doi.org/10.1016/j.jocn.2021.02.018>
 24. Kim H et al (2020) Deep learning-based method to differentiate neuromyelitis optica spectrum disorder from multiple sclerosis. *Front Neurol* 11:599042. <https://doi.org/10.3389/fneur.2020.599042>
 25. Liu Y et al (2019) Radiomics in multiple sclerosis and neuromyelitis optica spectrum disorder. *Eur Radiol* 29:4670–4677. <https://doi.org/10.1007/s00330-019-06026-w>
 26. Luo X et al (2022) Multi-lesion radiomics model for discrimination of relapsing-remitting multiple sclerosis and neuropsychiatric systemic lupus erythematosus. *Eur Radiol* 32: 5700. <https://doi.org/10.1007/s00330-022-08653-2>
 27. Rauschecker AM et al (2020) Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI. *Radiology* 295:626–637. <https://doi.org/10.1148/radiol.2020190283>
 28. Commowick O et al (2018) Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing

- infrastructure. *Sci Rep* 8:13650. <https://doi.org/10.1038/s41598-018-31911-7>
29. de Oliveira M et al (2022) Lesion volume quantification using two convolutional neural networks in MRIs of multiple sclerosis patients. *Diagnostics (Basel)* 12. <https://doi.org/10.3390/diagnostics12020230>
30. Gabr RE et al (2020) Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: a large-scale study. *Mult Scler* 26:1217–1226. <https://doi.org/10.1177/1352458519856843>
31. Griffanti L et al (2016) BIANCA (Brain Intensity AbNormality Classification Algorithm): a new tool for automated segmentation of white matter hyperintensities. *NeuroImage* 141: 191–205. <https://doi.org/10.1016/j.neuroimage.2016.07.018>
32. Hindholm AM et al (2021) Assessment of artificial intelligence automatic multiple sclerosis lesion delineation tool for clinical use. *Clin Neuroradiol* 32:643. <https://doi.org/10.1007/s00062-021-01089-z>
33. Schmidt P et al (2019) Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *NeuroImage Clin* 23:101849. <https://doi.org/10.1016/j.nicl.2019.101849>
34. Shiee N et al (2010) A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage* 49:1524–1535. <https://doi.org/10.1016/j.neuroimage.2009.09.005>
35. Zhang H et al (2021) ALL-Net: anatomical information lesion-wise loss function integrated into neural network for multiple sclerosis lesion segmentation. *NeuroImage Clin* 32:102854. <https://doi.org/10.1016/j.nicl.2021.102854>
36. Zhang Y et al (2022) A deep learning algorithm for white matter hyperintensity lesion detection and segmentation. *Neuroradiology* 64: 727–734. <https://doi.org/10.1007/s00234-021-02820-w>
37. Rakić M et al (2021) icobrain ms 5.1: combining unsupervised and supervised approaches for improving the detection of multiple sclerosis lesions. *NeuroImage Clin* 31:102707. <https://doi.org/10.1016/j.nicl.2021.102707>
38. Valverde S et al (2017) Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage* 155:159–168. <https://doi.org/10.1016/j.neuroimage.2017.04.034>
39. Carass A et al (2017) Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage* 148:77–102. <https://doi.org/10.1016/j.neuroimage.2016.12.064>
40. Weeda MM et al (2019) Comparing lesion segmentation methods in multiple sclerosis: input from one manually delineated subject is sufficient for accurate lesion segmentation. *NeuroImage Clin* 24:102074. <https://doi.org/10.1016/j.nicl.2019.102074>
41. Valverde S et al (2019) One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage Clin* 21:101638. <https://doi.org/10.1016/j.nicl.2018.101638>
42. Gasperini C et al (2019) Unraveling treatment response in multiple sclerosis: a clinical and MRI challenge. *Neurology* 92:180–192. <https://doi.org/10.1212/WNL.0000000000006810>
43. Cabezas M et al (2016) Improved automatic detection of new T2 lesions in multiple sclerosis using deformation fields. *AJNR Am J Neuroradiol* 37:1816–1823. <https://doi.org/10.3174/ajnr.A4829>
44. Sweeney EM, Shinohara RT, Shea CD, Reich DS, Crainiceanu CM (2013) Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. *AJNR Am J Neuroradiol* 34:68–73. <https://doi.org/10.3174/ajnr.A3172>
45. Krüger J et al (2020) Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks. *NeuroImage Clin* 28:102445. <https://doi.org/10.1016/j.nicl.2020.102445>
46. McKinley R et al (2020) Automatic detection of lesion load change in multiple sclerosis using convolutional neural networks with segmentation confidence. *NeuroImage Clin* 25:102104. <https://doi.org/10.1016/j.nicl.2019.102104>
47. Salem M et al (2018) A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. *NeuroImage Clin* 17:607–615. <https://doi.org/10.1016/j.nicl.2017.11.015>
48. Salem M et al (2020) A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. *NeuroImage Clin* 25: 102149. <https://doi.org/10.1016/j.nicl.2019.102149>
49. Rovira A et al (2022) Assessment of automatic decision-support systems for detecting active T2 lesions in multiple sclerosis patients. *Mult Scler* 28:1209. <https://doi.org/10.1177/13524585211061339>

50. Geurts JJ, Calabrese M, Fisher E, Rudick RA (2012) Measurement and clinical effect of grey matter pathology in multiple sclerosis. *Lancet Neurol* 11:1082–1092. [https://doi.org/10.1016/S1474-4422\(12\)70230-2](https://doi.org/10.1016/S1474-4422(12)70230-2)
51. Lucchinetti CF et al (2011) Inflammatory cortical demyelination in early multiple sclerosis. *N Engl J Med* 365:2188–2197. <https://doi.org/10.1056/NEJMoa1100648>
52. Thompson AJ et al (2018) Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol* 17:162–173. [https://doi.org/10.1016/S1474-4422\(17\)30470-2](https://doi.org/10.1016/S1474-4422(17)30470-2)
53. Geurts JJ et al (2005) Intracortical lesions in multiple sclerosis: improved detection with 3D double inversion-recovery MR imaging. *Radiology* 236:254–260. <https://doi.org/10.1148/radiol.2361040450>
54. Finck T et al (2020) Deep-learning generated synthetic double inversion recovery images improve multiple sclerosis lesion detection. *Investig Radiol* 55:318–323. <https://doi.org/10.1097/RLI.0000000000000640>
55. Bouman PM et al (2022) Artificial double inversion recovery images for (juxta)cortical lesion visualization in multiple sclerosis. *Mult Scler* 28:541–549. <https://doi.org/10.1177/13524585211029860>
56. Bouman PM, Steenwijk MD, Geurts JJG, Jonkman LE (2022) Artificial double inversion recovery images can substitute conventionally acquired images: an MRI-histology study. *Sci Rep* 12:2620. <https://doi.org/10.1038/s41598-022-06546-4>
57. Woolen SA et al (2020) Risk of nephrogenic systemic fibrosis in patients with stage 4 or 5 chronic kidney disease receiving a group II gadolinium-based contrast agent: a systematic review and meta-analysis. *JAMA Intern Med* 180:223–230. <https://doi.org/10.1001/jamainternmed.2019.5284>
58. Narayana PA et al (2020) Deep learning for predicting enhancing lesions in multiple sclerosis from noncontrast MRI. *Radiology* 294: 398–404. <https://doi.org/10.1148/radiol.2019191061>
59. Bodini B et al (2016) Dynamic imaging of individual remyelination profiles in multiple sclerosis. *Ann Neurol* 79:726–738. <https://doi.org/10.1002/ana.24620>
60. Wei W et al (2020) Predicting PET-derived myelin content from multisequence MRI for individual longitudinal analysis in multiple sclerosis. *NeuroImage* 223:117308. <https://doi.org/10.1016/j.neuroimage.2020.117308>
61. Sastre-Garriga J et al (2020) MAGNIMS consensus recommendations on the use of brain and spinal cord atrophy measures in clinical practice. *Nat Rev Neurol* 16:171–182. <https://doi.org/10.1038/s41582-020-0314-x>
62. Cole JH, Franke K (2017) Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends Neurosci* 40:681–690. <https://doi.org/10.1016/j.tins.2017.10.001>
63. Cole JH et al (2017) Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage* 163:115–124. <https://doi.org/10.1016/j.neuroimage.2017.07.059>
64. Franke K, Ziegler G, Kloppel S, Gaser C, Alzheimer's Disease Neuroimaging, I (2010) Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. *NeuroImage* 50:883–892. <https://doi.org/10.1016/j.neuroimage.2010.01.005>
65. Hogestol EA et al (2019) Cross-sectional and longitudinal MRI brain scans reveal accelerated brain aging in multiple sclerosis. *Front Neurol* 10:450. <https://doi.org/10.3389/fneur.2019.00450>
66. Kaufmann T et al (2019) Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nat Neurosci* 22: 1617–1623. <https://doi.org/10.1038/s41593-019-0471-7>
67. Cole JH et al (2020) Longitudinal assessment of multiple sclerosis with the brain-age paradigm. *Ann Neurol* 88:93–105. <https://doi.org/10.1002/ana.25746>
68. Fonteijn HM et al (2011) An event-based disease progression model and its application to familial Alzheimer's disease. *Inf Process Med Imaging* 22:748–759. https://doi.org/10.1007/978-3-642-22092-0_61
69. Fonteijn HM et al (2012) An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImage* 60:1880–1889. <https://doi.org/10.1016/j.neuroimage.2012.01.062>
70. Eshaghi A et al (2018) Progression of regional grey matter atrophy in multiple sclerosis. *Brain* 141:1665–1677. <https://doi.org/10.1093/brain/awy088>
71. Eshaghi A et al (2018) Deep gray matter volume loss drives disability worsening in multiple sclerosis. *Ann Neurol* 83:210–222. <https://doi.org/10.1002/ana.25145>
72. Tintore M et al (2015) Defining high, medium and low impact prognostic factors for developing multiple sclerosis. *Brain* 138:1863–1874. <https://doi.org/10.1093/brain/awv105>

73. Roca P et al (2020) Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. *Diagn Interv Imaging* 101:795–802. <https://doi.org/10.1016/j.diii.2020.05.009>
74. Zhao Y et al (2017) Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLoS One* 12:e0174866. <https://doi.org/10.1371/journal.pone.0174866>
75. Bendfeldt K et al (2019) MRI-based prediction of conversion from clinically isolated syndrome to clinically definite multiple sclerosis using SVM and lesion geometry. *Brain Imaging Behav* 13:1361–1374. <https://doi.org/10.1007/s11682-018-9942-9>
76. Wottschel V et al (2019) SVM recursive feature elimination analyses of structural brain MRI predicts near-term relapses in patients with clinically isolated syndromes suggestive of multiple sclerosis. *NeuroImage Clin* 24:102011. <https://doi.org/10.1016/j.nicl.2019.102011>
77. Zhang H et al (2019) Predicting conversion from clinically isolated syndrome to multiple sclerosis—an imaging-based machine learning approach. *NeuroImage Clin* 21:101593. <https://doi.org/10.1016/j.nicl.2018.11.003>
78. Pareto D et al (2022) Prognosis of a second clinical event from baseline MRI in patients with a CIS: a multicenter study using a machine learning approach. *Neuroradiology* 64:1383. <https://doi.org/10.1007/s00234-021-02885-7>
79. Young AL et al (2018) Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun* 9:4273. <https://doi.org/10.1038/s41467-018-05892-0>
80. Eshaghi A et al (2021) Author Correction: identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. *Nat Commun* 12:3169. <https://doi.org/10.1038/s41467-021-23538-6>
81. Eshaghi A et al (2021) Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. *Nat Commun* 12: 2078. <https://doi.org/10.1038/s41467-021-22265-2>
82. Pontillo G et al (2022) Stratification of multiple sclerosis patients using unsupervised machine learning: a single-visit MRI-driven approach. *Eur Radiol* 32:5382. <https://doi.org/10.1007/s00330-022-08610-z>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 29

Machine Learning for Cerebrovascular Disorders

Yannan Yu and David Yen-Ting Chen

Abstract

Cerebrovascular disease refers to a group of conditions that affect blood flow and the blood vessels in the brain. It is one of the leading causes of mortality and disability worldwide, imposing a significant socioeconomic burden to society. Research on cerebrovascular diseases has been rapidly progressing leading to improvement in the diagnosis and management of patients nowadays. Machine learning holds many promises for further improving clinical care of these disorders. In this chapter, we will briefly introduce general information regarding cerebrovascular disorders and summarize some of the most promising fields in which machine learning shall be valuable to improve research and patient care. More specifically, we will cover the following cerebrovascular disorders: stroke (both ischemic and hemorrhagic), cerebral microbleeds, cerebral vascular malformations, intracranial aneurysms, and cerebral small vessel disease (white matter hyperintensities, lacunes, perivascular spaces).

Key words Cerebrovascular disorders, Machine learning, Stroke, Cerebral microbleeds, Cerebral vascular malformations, Intracranial aneurysms, Cerebral small vessel disease, White matter hyperintensities, Lacunes, Perivascular spaces

1 Introduction

Cerebrovascular disorders are a group of conditions that affect blood vessels in the brain and cerebral blood circulation. Stroke is the most common presentation of cerebrovascular disorders. The majority of strokes are ischemic, caused by decreased blood flow to the brain leading to damage of brain tissue and neurologic dysfunction. Less common are hemorrhagic strokes, caused by blood extravasation out of cerebral blood vessels into the brain tissue itself (intracranial hemorrhage) or in spaces surrounding brain tissue (subarachnoid and subdural hemorrhage). Hemorrhagic strokes can lead to catastrophic injury due to increased intracranial pressure, decreased brain tissue perfusion, and damaged normal brain tissue. In 2019, there were 6.6 million deaths attributable to cerebrovascular disease worldwide; three million individuals died of ischemic stroke, 2.9 million died of intracerebral hemorrhage, and 0.4 million died of subarachnoid hemorrhage [1]. Stroke is the

second leading cause of death, accounting for 11.6% of all deaths globally, and the third leading cause of death and disability combined, contributing to 143 million disability-adjusted life years [2]. Cerebral small vessel disease encompasses a spectrum of disorders affecting the brain's small perforating arterioles, capillaries, and venules. It has a wide range of clinical manifestations, causing approximately 25% of strokes and contributing to approximately 45% of dementia cases [3]. Cerebral small vessel disease is highly prevalent in the elderly population, affecting from 5% of people at age 50 to almost 100% of people older than 90 years [3]. Intracranial aneurysms (IA) are due to ballooning in a blood vessel in the brain; if aneurysms rupture, they can lead to catastrophic subarachnoid hemorrhage with a mortality rate of 23–51% [4, 5] and permanent disability in 30–40% [4, 6]. Arteriovenous malformations (AVM) are due to a tangle of blood vessels in the brain that bypass normal brain tissue; AVMs can cause hemorrhage and seizures.

Cerebrovascular disorders are commonly diagnosed with imaging studies, and the treatment of some cerebrovascular disorders is based on imaging guidance. Common imaging modalities include computed tomography (CT), magnetic resonance imaging (MRI), and digital subtraction angiography (DSA). CT provides a rapid exam of brain tissue and brain vessels; some of the CT protocols will be mentioned in this chapter including non-contrast CT, CT angiography (CTA), and CT perfusion (CTP). Non-contrast CT is the exam of choice for diagnosing intracranial hemorrhage and also the exam of choice for initial triaging of ischemic stroke. However, ischemic stroke presentation on non-contrast CT depends mostly on stroke age, ranging from no change or subtle changes in 0–6 h to obvious hypoattenuation after 24 h. Post-contrast CT, depending on the detailed protocol, can highlight the vascular structures known as CTA often used in diagnose artery occlusion in ischemic stroke, IA, or AVM. Post-contrast CT can also calculate brain blood perfusion status known as CTP, commonly used in ischemic stroke triaging. MRI has various sequences that give tissue a particular appearance for medical diagnosis. Some of the sequences that will be mentioned in this chapter include the following. Diffusion-weighted imaging (DWI) measures water molecule movement restriction and is very sensitive to injured tissue in stroke. Perfusion-weighted imaging (PWI) and arterial spin labeling (ASL) both measure brain perfusion, but PWI requires contrast injection, while ASL does not. They are often used in ischemic stroke. Gradient-recalled echo (GRE) and susceptibility-weighted images (SWI) are both sensitive to iron and calcium deposition and used in blood product detection and can be used for detecting hemorrhage and small vessel disease. T2-weighted fluid-attenuated inversion recovery (FLAIR) is commonly used to detect stroke lesions >6 h and small vessel disease. For CT perfusion and MR

perfusion, quantitative perfusion maps can be calculated to estimate the blood perfusion status, common ones including cerebral blood flow (CBF), cerebral blood volume (CBV), time-to-maximum of residue function (Tmax), and mean transit time (MTT). DSA is a fluoroscopic technique (similar to X-ray) to visualize vasculature, which is used for the diagnosis and treatment of IA, ischemic stroke artery occlusions, and some AVMs.

Machine learning holds the promise of optimizing cerebrovascular disorder care, with the potential ability to improve or accelerate diagnosis and provide prognostication utilizing both clinical and imaging data.

2 Ischemic Stroke

Approximately 87% strokes are ischemic and 13% are hemorrhagic [1]. Ischemic stroke is due to reduced or absent blood supply to part of the brain, typically due to an occlusion or stenosis of a cerebral artery, leading to localized brain tissue damage and loss of neurological function. Ischemic damage to the brain is strongly time-dependent [7]. The only recommended treatments available to treat or mitigate damage due to ischemic stroke are IV thrombolysis within 4.5 h of symptom onset and endovascular thrombectomy within 24 h of symptom onset; these treatments are only approved for specific subsets of stroke patients [8]. Acute stroke therapies work to recanalize an occluded cerebral blood vessel and restore blood flow to ischemic or hypoperfused brain tissue, specifically via intravenous medication that can break up the occlusion (thrombolysis) or mechanical removal of the occlusion within the culprit artery (endovascular thrombectomy). Because clinical protocols are time-sensitive and standardized, timely diagnosis of ischemic stroke and rapid initiation of treatment are crucial steps in clinical practice [7]. Therefore, there is great potential for machine learning-based algorithms in acute ischemic stroke care. Figure 1 is a general example of how stroke is typically diagnosed and treated in the clinical setting.

In this section, we review studies that investigated machine learning application in large vessel occlusion (LVO) diagnosis, stroke onset time evaluation, stroke lesion segmentation, stroke outcome, and complication prediction. Common imaging modalities in acute stroke were computed tomography (CT) and magnetic resonance imaging (MRI) for stroke diagnosis and triaging and digital subtraction angiography (DSA) for both stroke diagnosis and treatment. Examples of those imaging modalities are demonstrated in Figs. 2 and 3.

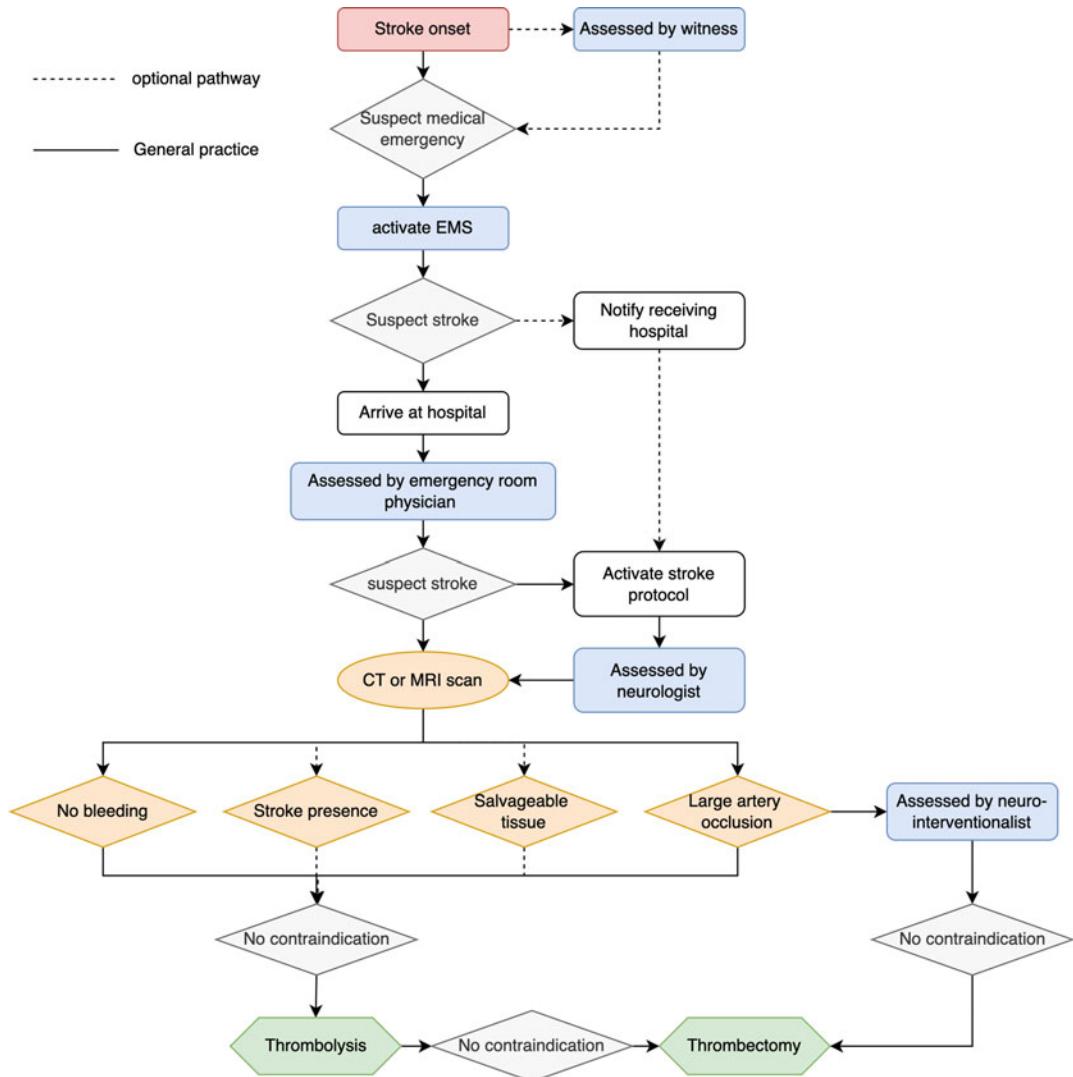


Fig. 1 General pathway of stroke diagnosis and treatment. Solid line represents general practice; dashed line represents optional pathway. EMS emergency medical service, CT computed tomography, MRI magnetic resonance imaging

2.1 Diagnosing Large Vessel Occlusion (LVO)

Large vessel occlusions are defined as blockages of the proximal intracranial arteries, accounting for approximately 24–46% of acute ischemic strokes [9]. Diagnosing an LVO is an important step of stroke diagnosis and treatment considerations; patients with LVO are potential candidates for endovascular thrombectomy, which is the most effective treatment available to recanalize an occluded artery [8, 10, 11]. Endovascular thrombectomy is a highly specialized procedure, and the personnel and equipment needed for thrombectomy are not widely available. Patients often need to be transferred from the hospital where they are initially evaluated to

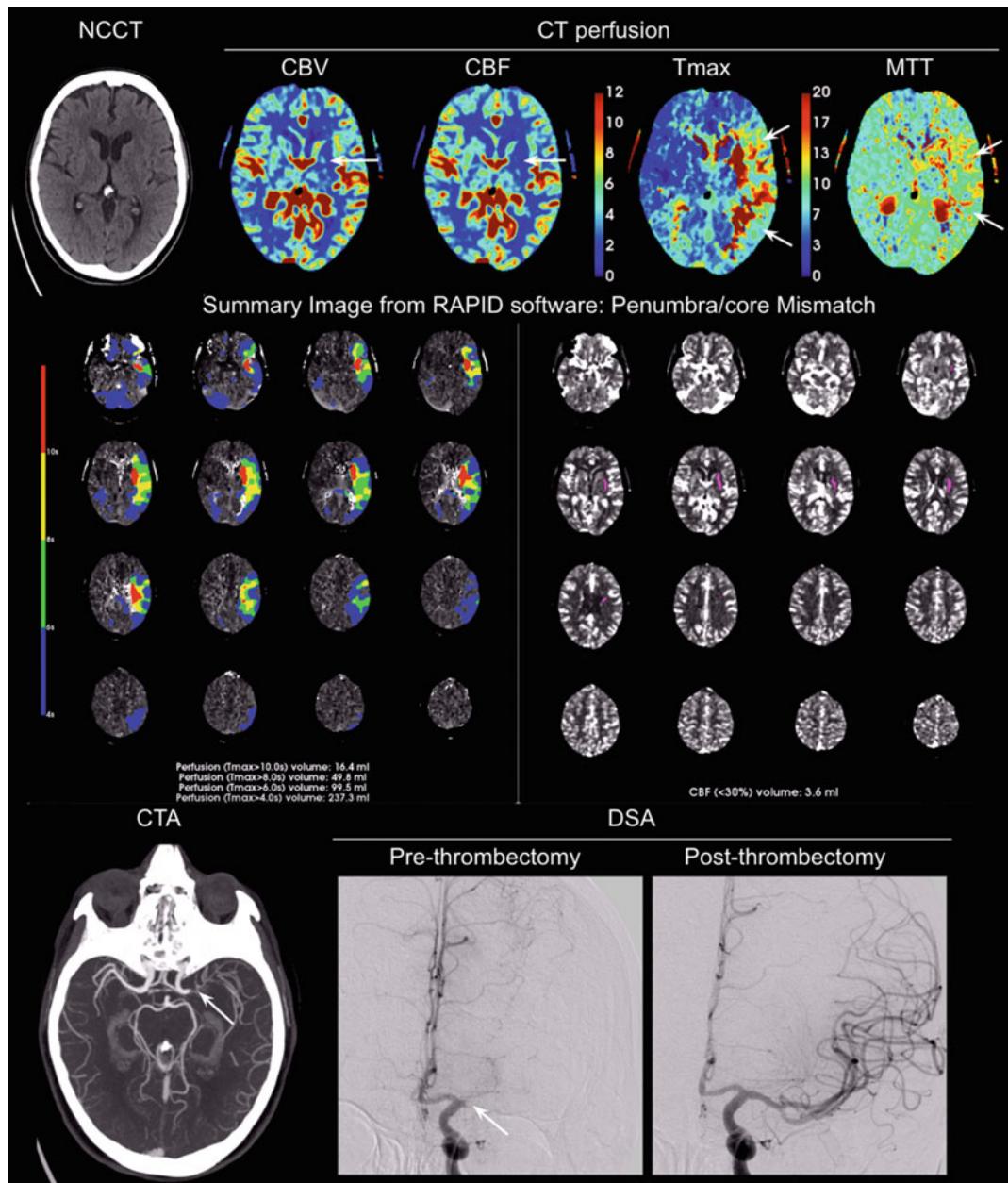


Fig. 2 Common CT scans used in acute stroke. This example case showed left-sided stroke (on the right side of the image) with occlusion of the middle cerebral artery (M1 segment). The NCCT had only very subtle changes, and the CT perfusion showed large perfusion deficit (asymmetrically low measures in CBF and asymmetrically high measures on Tmax and MTT) and small irreversible tissue injury. Penumbra/core mismatch is the volume ratio between prolonged Tmax area and decreased CBF area; the summary image from RAPID software showed mismatch ratio of 99.5 mL/3.6 mL. CTA showed middle cerebral artery main trunk (M1 segment) occlusion. The DSA image showed recanalization of the artery occlusion after thrombectomy. NCCT non-contrast computed tomography, CBV cerebral blood volume, CBF cerebral blood flow, Tmax time to maximum of the tissue residue function, MTT mean transit time, CTA computed tomography angiography, DSA digital subtraction angiography

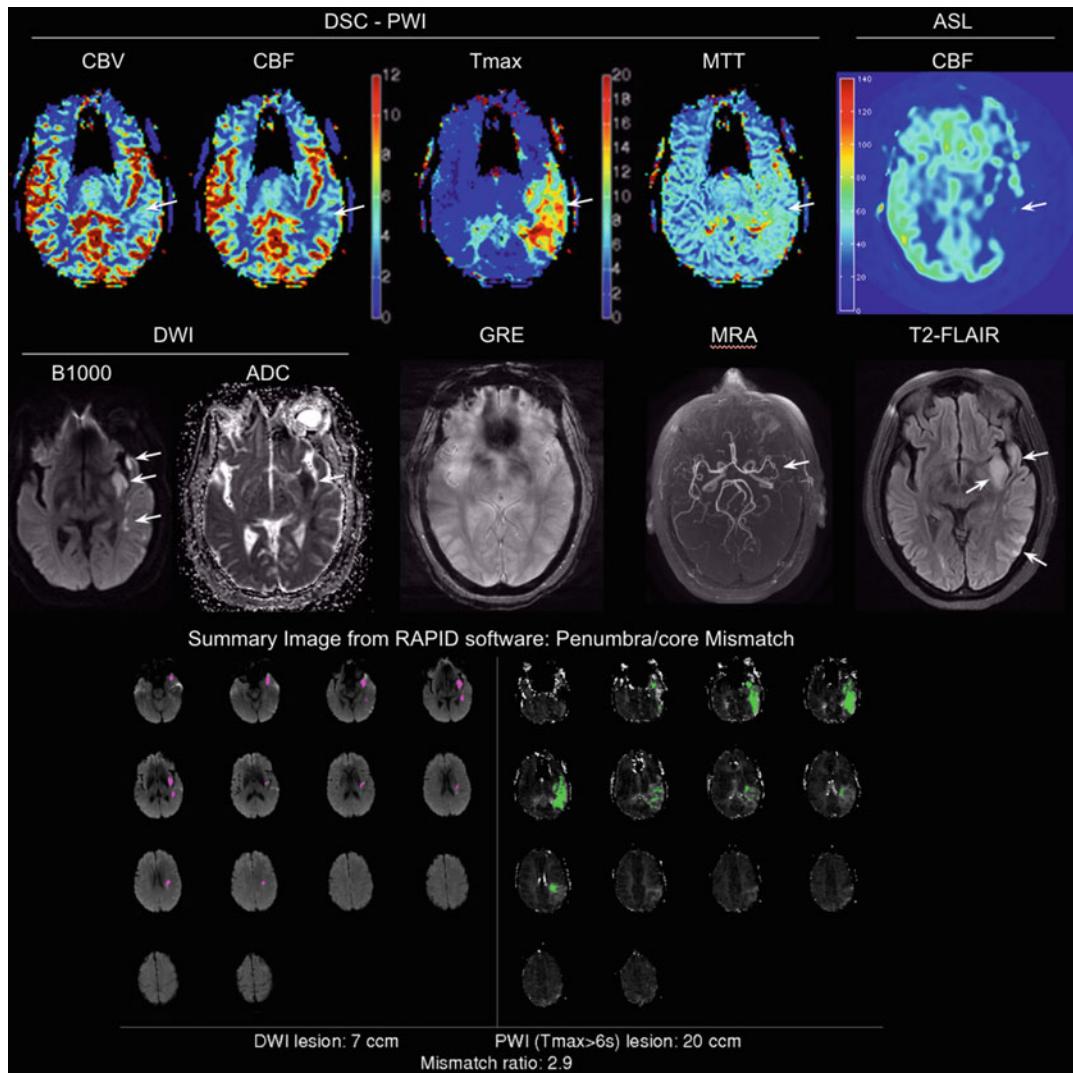


Fig. 3 Common MRI sequences used in acute stroke. This example case showed left-sided stroke (on the right side of the image) with occlusion of a middle cerebral artery branch (M2 segment, inferior division). The DSC-PWI and ASL showed the perfusion deficit (asymmetrically low measures in CBF and asymmetrically high measures on Tmax and MTT), which was much greater than the irreversible tissue injury on DWI, with a mismatch ratio of 2.9. GRE did not show blooming effect (a common finding of acute intra-arterial thrombus), MRA showed a left-sided large vessel occlusion (M2 segment, white arrow), and T2-FLAIR taken 24 h after the stroke showed injured brain tissue after the stroke (white arrows). DSC-PWI dynamic susceptibility contrast perfusion-weighted imaging. ASL arterial spin labeling, CBV cerebral blood volume, CBF cerebral blood flow, Tmax: time to maximum of the tissue residue function, MTT mean transit time, DWI diffusion-weighted imaging, ADC apparent diffusion coefficient, GRE gradient-recalled echo sequence, MRA magnetic resonance angiography, T2-FLAIR T2-weighted fluid-attenuated inversion recovery

a comprehensive stroke center with specialists who perform thrombectomy. Non-specialized hospitals must have the ability to reach the initial diagnosis of LVO-related stroke and arrange urgent

transfer to a comprehensive stroke center. During initial triage, automatic detection of LVO may accelerate the acute stroke protocol and patient transfer [12].

CT angiography (CTA) is the image modality of choice for rapid, non-invasive diagnosis of a large vessel occlusion. Several studies have used machine learning to demonstrate the feasibility to identify LVO on CTA. Viz.ai developed a commercial method of LVO detection that was achieved by a two-step analysis of CTA vessel segmentation via a 3D U-Net and large vessel classification via comparison of endpoint length and Hounsfield unit (a standardized unit for CT image pixel) value in MCA branch segmentation. Yahav-Dovrat et al. [13] reported the performance of this system in a prospective cohort of 404 stroke protocol CTAs. Seventy-two of the 404 stroke protocol CTAs had an LVO, and the software showed a sensitivity of 82%, a positive predictive value of 64%, and a negative predictive value of 96%. The relatively low sensitivity and positive predictive value may limit the clinical utility of the reported model, as the screening process of acute ischemic stroke requires high sensitivity. Stib et al. [14] trained convolutional neural networks (CNNs) with maximal intensity projection (MIP) images of multiphase CTA from 270 patients with LVO and 270 without LVO. The authors then tested the model in a balanced dataset of 62 patients, which showed a sensitivity of 100% and specificity of 77% by using all phases in multi-phase CTA, exceeding the performance of single-phase CTA with a sensitivity of 77% and specificity of 71%. To note, a non-deep learning-based commercial method from RAPID showed excellent sensitivity and specificity (above 95%) in an independent validation cohort [15]. These automated technologies have already become integrated into the clinical practice of many stroke systems of care, and further refinement of algorithm for center-specific population may improve the clinical performance.

LVO can also be detected from non-angiographic images, specifically non-contrast CT, which is more widely available than CTA. CTA requires intravenous contrast injection, which is typically not given to patients with kidney failure and/or an allergy to iodinated contrast. You et al. [16, 17] reported a XGBoost model trained with 200 cases' clinical data and non-contrast CT image features extracted from the bottleneck of U-Net; the model showed a sensitivity of 95.3% and specificity of 68.4% in 100 test cases. Olive-Gadea et al. [18] reported a DenseNet and decision tree-based prediction model to diagnose LVO from non-contrast CT images, showing a sensitivity of 83.1% and specificity of 85.1%, which exceeded the performance of a National Institutes of Health (NIH) stroke scale-based model.

Digital subtracted angiography (DSA) is an invasive diagnostic method for LVO used to guide interventional neuroradiologists treating the vascular occlusion. Thrombectomy treatment is

performed under the guidance of DSA to retrieve the thrombus causing occlusion. However, reading DSA images requires highly specialized training in interventional neuroradiology, and a real-time evaluation of treatment effect during the thrombectomy procedure is often required. Thrombolysis in cerebral infarction (TICI) scale is an evaluation on DSA for stroke treatment effect after thrombectomy procedure. Previous studies reported that the inter-reader agreement of TICI was low [19, 20]. Machine learning on DSA studies is challenging because the DSA contains 2D projection images from a 3D vasculature which are sensitive to the position of the X-ray detector plane, as well as temporal information that makes the data more similar to a video. When reading the DSA images, radiologists focus on the anatomical difference compared to the normal atlas, the speed of contrast filling into the arteries, the extent of contrast filling into the capillary system, and the contrast drainage from the veins.

Ueda et al. [21] collected DSA images with and without misregistration artifact and applied U-Net and convolutional patch generative adversarial network architecture as generator and discriminator networks to predict non-misregistered DSA from misregistered DSA. Zhang et al. [22] proposed a U-Net to track and segment the brain vessels from DSA, which could be the first step for building a diagnostic tool. As DSA is a 2D image with temporal information, studies used different strategies to blend these features into a neural network. Bhurwani et al. [23] proposed an ensembled convolutional neural network for post-thrombectomy DSA images and predict the reperfusion status. They achieved a sensitivity of 90% and specificity of 74% on diagnosing reperfusion after thrombectomy. Su et al. [24] proposed a curated algorithm including phase classification, motion correction, and perfusion segmentation to achieve final TICI scoring using ResNet-18. They achieved an agreement of 90% between the algorithm and human reader. To note, human-to-human agreement was 89%. Researchers from the same group [25] also designed a sophisticated network for spatial and temporal feature extraction and predict perforation, a complication from thrombectomy procedure. The model predicted perforation with precision of 0.83 and recall of 0.70, a performance similar to that of human expert readers.

In addition to classifying LVO on imaging, studies also showed it is feasible to predict LVO based on clinical evaluation, which could prepare the emergency medical services (EMS) for direct transport to comprehensive stroke centers [26–30]. Chen et al. [27] trained ANN models using tenfold cross-validation on 600 patients with 1:1 ratio of LVO and non-LVO using patients' NIHSS breakdown score, demographics, medical history, and risk factors as input. The ANN models reached sensitivity of 0.807 and specificity of 0.833. Wang et al. [26] from the same group then trained 8 machine learning models on 15,365 patients and test on

4215 patients using their NIHSS, demographics, medical history, and risk factors as input. They showed random forest model performed the best with an AUC of 0.831, sensitivity of 0.721, and specificity of 0.827.

2.2 Predicting Stroke Onset Time

In 14–27% of strokes, the symptom onset time is not known [31–33]. For those patients, identifying the likely onset time is crucial for proper treatment. Indeed, it is key to know if one is still within the treatment window for intravenous thrombolysis (within 4.5 h) or endovascular therapy (within 6 h if presence of LVO plus no extensive lesion on non-contrast CT or 24 h if presence of LVO and target mismatch on perfusion imaging). MRI plays a key role in estimating the duration of stroke. Studies have shown that fluid-attenuated inversion recovery (FLAIR) usually detects ischemic lesion after 3–6 h of stroke onset [34, 35], in contrast to diffusion-weighted imaging (DWI), which detects ischemic lesions within minutes of stroke. Therefore, the “mismatch” between FLAIR and DWI may be used as a clock for determining stroke onset time [36]. Lee et al. [37] captured 89 vector features from DWI and FLAIR imaging and trained machine learning models including logistic regression, support vector machine, and random forest to classify if the stroke onset is within 4.5 h. They found the machine learning models were more sensitive (75.8% vs 48.5%, $p = 0.01$) but less specific (82.6% vs 91.3%, $p = 0.15$) compared to human readers. Similar results were also achieved by other research groups [38]. Perfusion MRI has not been studied in the past for determining the stroke onset time. Ho et al. [39, 40] extracted deep features using an autoencoder from perfusion MRI to classify whether stroke onset time was within 4.5 h (the current time window for intravenous tissue plasminogen activator [tPA]). Using input DWI, apparent diffusion coefficient (ADC), FLAIR, and perfusion-weighted images, they achieved a ROC AUC of 0.765. This approach outperformed DWI-FLAIR-based machine learning methods (AUC of 0.669) and clinical methods (AUC of 0.58) in the same dataset. The use of imaging to determine the time of stroke onset may increase the number of patients eligible for time-limited stroke treatments, such as intravenous thrombolysis [31].

2.3 Stroke Lesion Segmentation

Non-contrast-enhanced CT scan is the most common initial imaging obtained for stroke patients. Therefore, CT datasets are usually much more common and larger than MRI datasets. However, it is generally more challenging to diagnose early stroke or predict final stroke lesions on CT than MRI, as changes on CT related to early hyperacute phase (<6 h) of ischemic stroke are very subtle, including loss of gray and white matter differentiation, hypoattenuation of deep nuclei, and cortical hypodensity with associated parenchymal swelling and gyral effacement. The Alberta Stroke Program

Early CT Score (ASPECTS) is a scoring system that assesses stroke lesion presence based on early hyperacute phase changes on non-contrast CT image; scores range from 0 to 10, with 0 representing extensive ischemic damage and 10 representing no evidence of ischemia [41]. Current guidelines recommend reperfusion treatment for those with high ASPECTS [8], meaning less injured tissue, but ongoing research and trials are investigating the benefit of treating low ASPECTS stroke patients [42]. DWI/ADC is the most common and accurate MRI sequence to identify early stroke lesions (using a threshold of $ADC \leq 620 \times 10^{-6} \text{ mm}^2/\text{s}$). In addition, automated segmentation on MRI/CT would benefit acute treatment decisions as well as enable researchers to conduct clinical research on a much larger scale.

Many studies have showed the use of machine learning for stroke lesion segmentation on acute to subacute CTs and MRIs [43–57]. Kuang et al. [58] trained a random forest classifier on non-contrast CT images from 157 stroke patients to predict the ASPECTS score on MRI scanned within 1 h after the CT image and tested on 100 patients. They achieved a sensitivity of 66.2% and specificity of 91.8% in 100×10 ASPECTS regions and sensitivity of 97.8% and specificity of 80% in classifying ASPECT >4 and ≤ 4 . Qiu et al. [57] from the same group used the same dataset to segment the early stroke lesion on non-contrast CT images using MRI as ground truth. They proposed a random forest algorithm with sophisticated feature engineering of distance feature, atlas encoded lesion location feature, and U-Net generated probability map of lesions from a separate dataset as input. They showed good correlation between predicted stroke lesion volume and ground truth ($r = 0.76$) and mean volume difference of 11 mL. Two commercial software programs for automatic ASPECTS scoring (e-ASPECTS, Brainomix, and Rapid ASPECTS, iSchemaView) are available and reported to be not inferior or even more accurate than clinicians [59–64].

The Ischemic Stroke Lesion Segmentation (ISLES) 2015 challenge provided training and testing data for subacute stroke lesion segmentation using MRI sequences including DWI and FLAIR. In this challenge, the highest performance for lesion segmentation was achieved by a 3D CNN with Dice score coefficient (DSC) of 0.57 [45]. Chen et al. [43] developed a two-step method to segment stroke lesions from DWI, reaching a DSC of 0.67. The first step was using an encoder-decoder CNN to propose a lesion segmentation, with a second step CNN which took patches of original DW images and previous output at multiple scales as input and classified the proposed segmentation as true or false. Other studies reached similar results (DSC 0.64–0.76) with 2D and 3D encoder-decoder CNNs [47–50]. The ISLES 2018 challenge provided training and testing data for acute stroke CT perfusion imaging to predict irreversibly injured tissue defined on DWI [65]. The top team

used a 3D multi-scale U-Net with atrous convolution algorithm and achieved an average DSC and an average absolute volume difference of 0.51 and 10.2 mL, respectively [66]. Other studies also reached similar results but were less accurate than the top performing team (DSC 0.44–0.49) [67, 68].

The aforementioned methods require manual labeling of stroke lesions on many images to serve as training, which is expensive and limits the scale of medical image deep learning research. For this reason, Zhao et al. [52] explored semi-supervised algorithms (a combination of K-means clustering and CNN) in a weakly labeled stroke segmentation dataset using acute DWI and ADC, reaching a mean DSC of 0.64. Federau et al. [53] explored 3D U-Net segmentation using a dataset augmented with synthetic stroke lesions on DWI, achieving a DSC of 0.72. More recently, Zhang et al. [51] utilized a feature pyramidal network [69] and a U-Net with multi-plane (axial, sagittal, and coronal planes) DWI to perform lesion segmentation, which achieved a DSC of 0.62. As radiologists usually interpret MRI by looking at different sequences, neural networks that take different imaging sequences as input and “fuse” their information are an important research direction to improve the diagnosis.

Winzeck et al. [55] proposed to train an ensemble of CNNs instead of individual CNNs. The authors adopted the CNN structure from the highest performance model in ISLES 2015 challenge. They found that an ensemble of five 3D CNNs segmented the DWI lesion from ADC, DWI, and B0 images more accurately than individual CNNs (median DSC 0.82 vs 0.79). Wu et al. [44], from the same group, trained the ensemble of CNNs with a multi-center, multi-vendor dataset with ADC, DWI, and B0 data and found that it performed better than models trained with a single-center dataset, with a median DSC of 0.86 (IQR 0.79–0.89). Although the model performance cannot be directly compared between papers as they all used different test datasets, this chapter has reported the highest DSC in stroke lesion segmentation so far.

2.4 Predicting Stroke Lesions in the Future

As compared to the stroke lesion segmentation on a single-time point imaging, segmenting a final lesion or hemorrhagic transformation on follow-up images using baseline CT/MRI is a way to predict patient clinical and radiographic outcome in the future. In particular, methods that can predict individual response to treatment (e.g., predicting the future outcomes in the presence and absence of treatment) can be useful to determine whether the treatment would benefit to this individual.

The ISLES challenges from 2016 and 2017 were focused on stroke lesion prediction from initial MRIs, including diffusion and perfusion imaging [70]. Compared to human inter-reader agreement of DSC of 0.58, the best performing model, using an

encoder-decoder CNN, achieved a DSC of 0.32 [70]. Using data from this challenge, Pinto et al. [71] proposed an encoder-decoder CNN combined with 2D gated recurrent unit layers [72], with the TICI score fused at the end to generate lesion predictions based on different TICI scores. The model had a similar DSC of 0.35. Nielsen et al. [73] used a CNN to predict the final stroke lesion using baseline DWI and MR perfusion and reported an ROC AUC of 0.88. They also found CNNs trained with either treatment or no treatment predicted different stroke lesions, suggesting a role to use such models to explore differential outcomes with therapy. Ho et al. [74] proposed a CNN model to predict lesions directly from PWI source images (i.e., rather than from the parameter maps created by post-processing software), which reached a similar ROC AUC of 0.871. Yu et al. [75] showed that an attention-gated U-Net model could predict final stroke lesions at 2–7 days from baseline MR perfusion and diffusion images regardless of reperfusion status with a median DSC of 0.53 and ROC AUC of 0.92. In a separate study aimed at providing more accurate penumbra and ischemic core information, Yu et al. [76] pre-trained an attention-gated U-Net model with DWI and MR perfusion maps in patients with partial reperfusion or unknown reperfusion and then fine-tuned this pre-trained model with minimal reperfusers to predict penumbra and major reperfusers to predict ischemic core. The model achieved a median DSC of 0.60 for penumbra and 0.57 for ischemic core, exceeding the performance of the automated penumbra and ischemic core segmentation from state-of-the-art software. In a slightly different approach, Wang et al. [77] used a CNN to identify penumbral tissue (as defined by the Tmax perfusion parameter from contrast PWI) on non-contrast arterial spin labeling (ASL) with an ROC AUC of 0.958 which provided similar stroke triaging in 92% of cases without the need to inject a contrast agent.

It is more challenging to predict the final stroke lesion from CT image as the markers are not correlated with tissue injury as well as DWI. Robben et al. [78] proposed a CNN with parallel inputs from source CT perfusion images and clinical metadata, which achieved a mean DSC of 0.48. An ablation study was also performed, which showed in addition to image information, time from imaging to treatment also influenced the model prediction. Amador et al. [79] applied temporal CNN to predict the final lesion from the baseline CT perfusion source image, which achieved a DSC of 0.33. Kuang et al. [80] trained a random forest model from 67 patients' CT perfusion maps and clinical data and tested in 137 patients. They found the model reached a median volumetric difference of –3.2 mL and DSC of 0.388 and the model was significantly more accurate than thresholding methods (Tmax thresholding and CBF thresholding), although the reperfusion status of those patients were heterogeneous.

2.5 Predict Hemorrhagic Transformation

Hemorrhagic transformation is a potential complication of stroke treatment. Large hemorrhagic transformation can be lethal. Predicting hemorrhagic transformation after reperfusion therapy has been investigated in the past using statistical methods. To improve the prediction, Yu et al. [81, 82] proposed a long short-term memory network (LSTM) to predict the segmentation of hemorrhagic transformation lesion identified by gradient-recalled echo (GRE) sequence performed at 24 h after stroke onset, using baseline MR perfusion as input. The model demonstrated an ROC AUC of 0.894, which was higher than a previous SVM approach (ROC AUC of 0.837). Jiang et al. [83] included multi-parametric MRI and clinical data to predict the presence of hemorrhagic transformation. The image sequences were separately fed in to inception V3 architecture and connected with clinical data at the fully connected layers. The model achieved a high AUC of 0.932 and an accuracy of 0.873 in binary classification of hemorrhagic transformation.

2.6 Predicting Stroke Clinical Outcomes

Compared to predicting future stroke lesions on images, clinical outcome prediction is more difficult for several reasons. The most common scoring system, the modified Rankin score (mRS), is nonlinear and subjective, and the unit of analysis is each patient rather than each voxel (Table 1). The majority of the previously published studies used non-imaging data as input to predict clinical outcomes using simple statistical or more complex machine learning models [84–89]. However, images may provide more information such as the spatial location of infarct and hemorrhage and the presence of brain atrophy. Osama et al. [90] proposed a parallel multi-parametric feature-embedded Siamese neural network [91] to classify 3-month mRS from 0 to 4 using the MRI perfusion maps and clinical data from the ISLES 2017 challenge. This model achieved an average accuracy of 37% on each class using leave-one-out cross-validation testing. Nishi et al. [92] proposed a U-Net with DWI as input and stroke lesion segmentation as output. Then the bottleneck features of the U-Net were extracted to predict whether the 3-month mRS would be greater than 2, a common metric of good clinical outcome. This method achieved a ROC AUC of 0.81, exceeding the performance of ASPECTS Score (ROC AUC of 0.63) and ischemic core volume models (ROC AUC of 0.64). These studies show promise that automated imaging analysis might be helpful in the prediction of clinical outcomes, but further study into these complex and ambitious predictions is needed.

Table 1
Modified Rankin scale

0	No symptoms at all
1	No significant disability despite symptoms; able to carry out all usual duties and activities
2	Slight disability; unable to carry out all previous activities, but able to look after own affairs without assistance
3	Moderate disability; requiring some help, but able to walk without assistance
4	Moderately severe disability; unable to walk without assistance and unable to attend to own bodily needs without assistance
5	Severe disability; bedridden, incontinent, and requiring constant nursing care and attention
6	Dead

**2.7 Predicting
Cerebral Blood Flow
(CBF) and
Cerebrovascular
Reserve (CVR)**

Sometimes, it is useful to obtain more accurate images of biomarkers that drive stroke severity, such as CBF. The current CBF gold standard, O-15 water positron emission tomography (PET), is much less accessible than MRI or CT given its strict requirement for radiotracer production within the facility and exposure to radiation. ASL, a non-invasive MRI sequence measuring CBF without the use of intravenous contrast, allows repeat examination and limits any potential adverse effects from contrast or radiotracer agent. Although ASL has been improved over the last decades, it has low sensitivity, frequently underestimates CBF in areas with delayed collateral flow, and is prone to a range of artifacts. Guo et al. investigated whether a U-Net CNN can produce PET-like CBF maps from ASL and structural images [93]. Compared to the ASL CBF, the synthetic PET CBF map derived from the ASL and structural MRI scans had a significantly higher structural similarity index (0.854 ± 0.036 vs 0.743 ± 0.045). By training on both normal subjects and patients with cerebrovascular disease, they showed similar good performance to predict a PET CBF map regardless of disease status.

CVR is measured by calculating relative CBF change ($r\Delta\text{CBF}$) before and after a vasodilating drug. Patients with low CVR are at higher risk of future stroke, and the identification of these patients may be helpful in the initiation of preventative treatments, such as aggressive medical therapy, carotid endarterectomy, or carotid stent placement [94]. Acetazolamide, a carbonic anhydride inhibitor, is typically used as a vasodilator to measure CVR. It is generally safe, but it is contraindicated in patients with sulfa allergies or severe kidney and liver diseases. Some patients may present with stroke-like symptoms during the test. These symptoms, although transient and rare, may unsettle patients and medical staff.

To further simplify the measurement of CVR, Chen et al. [95] investigated the feasibility of a drug-free CVR measurement using a

U-Net CNN based on the work of Guo et al. [93]. The study also investigated several input combinations (MRI + PET vs MRI only) to determine whether baseline O-15 PET CBF information is required. Using a ground truth of O-15 PET rΔCBF in a cohort of Moyamoya disease patients (a condition with chronic narrowing of brain arteries leading to increased stroke risk), they showed that using the baseline MRI alone resulted in better performance at predicting regions with compromised CVR than the current clinical method using ASL before and after acetazolamide injection. Such a method may find use in estimating CVR from routine MRI scans acquired as part of clinical practice, obviating the need for either PET or acetazolamide.

3 Hemorrhagic Stroke or Intracranial Hemorrhage

Hemorrhagic stroke, also known as intracranial hemorrhage, accounts for approximately 13% of all strokes. Hemorrhagic stroke was found to have similar total death (three million yearly) and disability (69 million disability-adjusted life year) than ischemic stroke, although the incidence of ischemic stroke was twice as great [96]. Hemorrhagic stroke is commonly diagnosed through non-contrast CT or MRI (GRE or SWI are particularly sensitive to hemorrhage). Important considerations on the diagnosis and triaging include the presence, location, volume, and expansion of the hemorrhage. Chilamkurthy et al. [97] trained a ResNet with a large dataset of 300,000 CT scans to detect critical findings on CT including hemorrhage. The model was tested on 500 CT scans with high AUCs for detecting hemorrhage. However, the performance was not as good as expert radiologists. Lee et al. [98] proposed an ImageNet pre-trained deep CNN that was further trained on 904 CT cases of acute intracranial hemorrhage to detect hemorrhage and classify the 5 subtypes of hemorrhage. They tested in independent test datasets with about 400 cases and found the model achieved similar performance to expert radiologists with a sensitivity of 92–98% and specificity of 95%. In addition, the researchers attempted to explain this CNN model using the attention map, which showed that the model had a similar process that mimics the radiologists' workflow. Kuo et al. [99] trained a CNN with over 4000 head CT scans to classify and segment intracranial hemorrhages. They showed the model achieved an AUC of 0.991 on 200-case independent test set, with good performance in case with very small and subtle hemorrhagic lesions.

Machine learning has also been applied to diagnose the etiology of intracranial hemorrhage, examples including microbleeds, vascular malformation, and intracranial aneurysms. These topics are reviewed in separate sections.

4 Cerebral Vascular Malformation

Cerebral vascular malformations occur in 0.1–4.0% of the general population. Arteriovenous malformations (AVMs) are the most dangerous cerebral vascular malformation and can cause hemorrhage, seizures, headaches, and focal neurologic deficits.

Identifying intraparenchymal hemorrhage caused by AVMs on non-contrast-enhanced CT could be useful in triaging patients to appropriate treatment. Zhang et al. [100] selected radiomic features from 11 filter-based feature selection methods and applied multiple supervised machine learning algorithms to classify the intraparenchymal hemorrhage as AVM-related or other etiology. The best model was AdaBoost classifier, which achieved an AUC of 0.957, a sensitivity of 88.9%, and a specificity of 93.7% in the test set.

Stereotactic radiosurgery is most successful when used to treat small AVMs (diameter <3 cm) or in deep and eloquent areas that would engender great neurologic risk with attempted resection. Its performance relies on the accuracy of delineating the target AVM, since partial volume irradiation may result in obliteration failure and remained symptoms. Recently, Wang et al. [101] proposed a three-dimensional V-Net to automatically segment the AVMs on contrast CT images to guide stereotactic radiosurgery. They compared the V-Net model performance with human readers and achieved an average DSC of 0.85 and an average volume error of 0.076 mL among 80 patients.

Adverse radiation effects after stereotactic radiosurgery include cyst formation which may require surgical intervention and radiation-induced changes which may lead to permanent neurological deficits in 1–3% of the patients. Deep AVMs (located in the thalamus, basal ganglia, and brainstem), large AVMs, large radiation treatment volume, and repeated radiosurgery are risk factors to develop neurologic deficits after radiosurgery. Lee et al. [102] proposed an unsupervised classification with fuzzy c-means clustering to analyze the AVM nidus on T2-weighted MRI and analyzed the association between brain parenchyma component near the nidus and radiation-induced changes. The model automatically segmented nidus, brain parenchyma, and cerebrospinal fluid components in the radiation-exposed region. Compared with manual segmentation, the proposed algorithm achieved a DSC of 0.795. The automatically segmented brain parenchyma was associated with radiation-induced changes.

5 Intracranial Aneurysms

Intracranial aneurysms (IAs) have a prevalence of 3.2% in the general population [103, 104]. IA rupture accounts for 80–90% of spontaneous subarachnoid hemorrhages [5, 105], which is usually a catastrophic event, with a mortality rate of 23–51% [4, 5] and permanent disability in 30–40% [4, 6]. Survivors often suffer from long-term neuropsychological deficits and decreased quality of life. Although DSA is the gold standard to diagnose an aneurysm, unruptured IAs can be detected with non-invasive imaging techniques such as MR angiography (MRA) or CT angiography (CTA). Early diagnosis of IAs can benefit from clinical management which may prevent their rupture [106, 107]. However, there are two unmet clinical needs for IA: diagnosis and management.

5.1 Difficulty in Aneurysm Detection

Because of the small size of IAs and the complexity of intracranial vessels, aneurysm detection can be time-consuming and requires subspecialty training. It renders two challenges. First, there is a suboptimal inter-observer agreement ($\kappa = 0.67\text{--}0.73$) in the detection of IA from CTA and MRA [108]. The interpretation may vary depending on the level of expertise. Therefore, the sensitivity of detecting IA in CTA and MRA can range from 60% for a resident to 80% for a neuroradiologist [109]. Second, there is a high false-negative rate in detecting small aneurysms with diameter less than 5 mm. It has been reported that the sensitivity of detecting IAs of less than 5 mm is 57–70% [108, 110] for CTA and 35–58% for MRA [109, 110]. In comparison, the sensitivity of detecting IAs larger than 5 mm is 94% and 86% for CTA and MRA. Given all the difficulties mentioned above, there is a clinical need to have high-performance computer-assisted diagnosis (CAD) tools to aid in detection, increase efficiency, and reduce disagreement among observers which may potentially improve the clinical care of patients.

5.1.1 AI Algorithm for Intracranial Aneurysm Detection

There have been several studies showing that CAD program can automatically detect IA in MRA or CTA. The conventional CAD systems, based on manually designed imaging features, such as vessel curvature, thresholding, or a region-growing algorithm, have shown good performance in detecting IA [111, 112]. However, these conventional methods were developed on very small datasets and had to be modified manually when applied to new images. New deep learning-based methods directly learn the most predictive features from a large dataset of labeled images. They have better performance and greater generalizability than conventional methods. Deep learning has also been used for IA detection in MRA and CTA, and several studies have shown decent results [113–116].

The diagnostic accuracy of models using various imaging modalities has been studied. Digital subtraction angiography (DSA), an invasive vascular imaging procedure, is the gold standard to diagnose an aneurysm. Zeng et al. [117] applied 2D CNN on 3D DSA by concatenating five consecutive rotational angles of the DSA image patch as model input. The model reached an accuracy of 99%. Duan et al. [118] performed a similar task but on 2D DSA. It is more difficult due to less identifiable features in the 2D projection image, especially the differentiation between the vessel overlaps and an aneurysm. They proposed a two-stage detection system: First, the neural network localized the target region on the DSA using feature pyramid network. Second, the anchor box of aneurysm and vessel overlaps was generated by dual input of anterior-posterior view and lateral view into another feature pyramid network. The model reached an AUC of 93.5%.

MRA and CTA offer non-invasive diagnosis of intracranial aneurysms. Nakao et al. [113] and Sichtermann et al. [119] showed the feasibility of using CNN for aneurysm detection on time-of-flight (TOF) MRA. More recently, Ueda et al. [114] trained a ResNet-18 model to detect aneurysms on using 683 TOF MRAs. The model was tested on both internal data and external data with sensitivity and specificity above 90%. Park et al. [115] proposed a 3D CNN with an encoder-decoder structure to segment the intracranial aneurysms from CT angiography. Similar to U-Net, the model contains skip connections to transmit output directly from the encoder to the decoder. The encoder was pre-trained using videos labeled with human actions. The model was trained, validated, and tested using 611, 92, and 115 CTAs. Augmenting physicians with artificial intelligence-produced segmentation resulted in improvement in sensitivity, accuracy, and interrater agreement when compared with no augmentation. Faron et al. [120] showed similar results in 3D TOF MRA with a smaller dataset.

5.2 Difficulties in Aneurysm Risk Evaluation

Once an IA is detected in imaging study, clinicians must determine how to manage an unruptured IA. Overall, IAs have a low annual rupture risk of 0.95% [121]. Current treatments to prevent IA rupture include open neurosurgical clipping or endovascular embolization; both have a relatively high peri-operative risk of stroke and death (3–10%) [122]. Therefore, the management of unruptured aneurysm remains controversial [123]. Currently, the decision on whether to intervene is mainly based on aneurysm size. If an IA is larger than 5 mm in diameter in the anterior cerebral circulation or larger than 7 mm in the posterior circulation, surgical treatment is considered [123]. If an IA is smaller than these thresholds, follow-up observation with serial imaging is typically pursued [124]. Change in size of an IA during the follow-up period is a warning sign of impending rupture and often leads to surgical or

endovascular treatment. However, IA rupture depends on multiple factors in addition to size, including aneurysm shape and location as well as hemodynamics of the aneurysm, blood pressure, and mental and physical stress of the patient [121, 125]. It is not optimal to make the decision to intervene solely on size criteria, given risk of rupture is multifactorial. Moreover, follow-up serial imaging takes time, and rupture may occur during the observation period [126–128].

5.2.1 AI-Based Aneurysm Risk Prediction Model

A more comprehensive morphological evaluation of IA would be optimal; it ideally would include data on aneurysm shape, geometry, presence of a daughter sac, volume, and comparison of IA morphology across serial scans. Deep learning-based methods have the potential to automatically perform precise IA segmentation and provide efficient tools for the morphological evaluation of IA. Furthermore, machine learning methods can take high-dimensional, cross-domain inputs and directly learn from the labeled data to construct sophisticated prediction models. Feature ranks derived from the machine learning model could provide (? information) on individual factors that can influence model prediction.

Several studies have attempted to segment aneurysms using deep learning [129, 130]. Podgorsak et al. [130] used a CNN with encoder and decoder architecture to segment aneurysms on DSA, achieving a DSC above 0.9 for intracranial aneurysms.

Optimization of treatment decisions for unruptured small aneurysms [and patients with multiple aneurysms] is needed. Studies have applied machine learning algorithms to predict the outcomes of unruptured aneurysms [131–137]. Liu et al. [132] used morphologic features derived from DSA and machine learning models to predict if an aneurysm was unstable (defined as rupture within 1 month), aneurysm growth, and symptomatic aneurysms. They found that aneurysms with a diameter between 4 and 8 mm and irregular morphology indicate the aneurysm instability with an area under curve (AUC) of 0.85 in a separate test set. Similarly, Kim et al. [133] used CNN on small aneurysms based upon rotational DSA and showed that the model had better performance on the prediction of aneurysm rupture than human predictions.

Tanioka et al. used machine learning-based methods with morphological and hemodynamic parameters as inputs to achieve relatively high accuracy (71.2–78.3%) in predicting rupture status of IA [138]. They found projection ratio, irregular shape, and size ratio were important for the discrimination of ruptured aneurysms. Shi et al. further included clinical data to morphologic and hemodynamic information, to construct a machine learning model to predict IA rupture and reported areas under the curve of 0.88–0.91 [139].

After aneurysm rupture, predicting common complications of aneurysmal subarachnoid hemorrhage such as vasospasm, delayed cerebral ischemia, and functional outcome could help guide patient care. Kim et al. [140] used clinical factors and morphological features of an aneurysm to predict vasospasm after IA rupture with a random forest regressor. The model achieved an accuracy rate of 0.855 (AUC of 0.88). Ramos et al. [141] used clinical and CT image features to predict delayed cerebral ischemia using multiple machine learning algorithms. The best model reached an AUC of 0.74. Similarly, Rubbert et al. [142] used clinical and imaging features to predict 6-month dichotomized modified Rankin scale using random forest, with an accuracy of 71%.

6 Cerebral Small Vessel Disease

Cerebral small vessel disease (cSVD) encompasses a spectrum of disorders affecting the brain's small perforating arterioles, capillaries, and probably venules [143], which cause various focal and global brain lesions that can be detected on pathological examination and brain imaging [144]. cSVD has a wide range of clinical manifestations. Although many affected patients may remain asymptomatic, cSVD may herald patients at risk for acute ischemic stroke or intracerebral hemorrhage; it can also present as an insidious clinical course associated with progressive cognitive decline, development of mood disorders, and gait disturbance [145]. cSVD causes about one-fourth of all acute ischemic strokes and is a major risk factor for hemorrhagic strokes [146–148]. It is the most common cause of vascular dementia and mixed dementia, which often occurs with Alzheimer's disease, and contributes to about one-half of all dementias worldwide, thus causing a massive health burden [146, 149, 150].

6.1 Imaging Features of cSVD

Neuroimaging plays a pivotal role in the diagnosis and evaluation of cSVD [143]. According to the STAndards for ReportIng Vascular changes on nEuroimaging (STRIVE), the imaging features of cSVD include recent small subcortical infarcts, white matter hyperintensities (WMH) of presumed vascular origin, lacunes, enlarged perivascular spaces (PVS), and cerebral microbleeds (CMBs) (Fig. 4) [144]. These imaging findings, either individually or in combination, are associated with cognitive impairment, dementia, depression, mobility problems, increased risk of stroke, and worse outcomes after stroke [146, 151–153]. The quantification of cSVD imaging features is important for disease severity evaluation and clinical prognostication [154, 155]. However, these lesions are generally small and widespread in the brain, rendering manual inspection and segmentation laborious and prone to error. Machine learning algorithms have great potential in the automatic

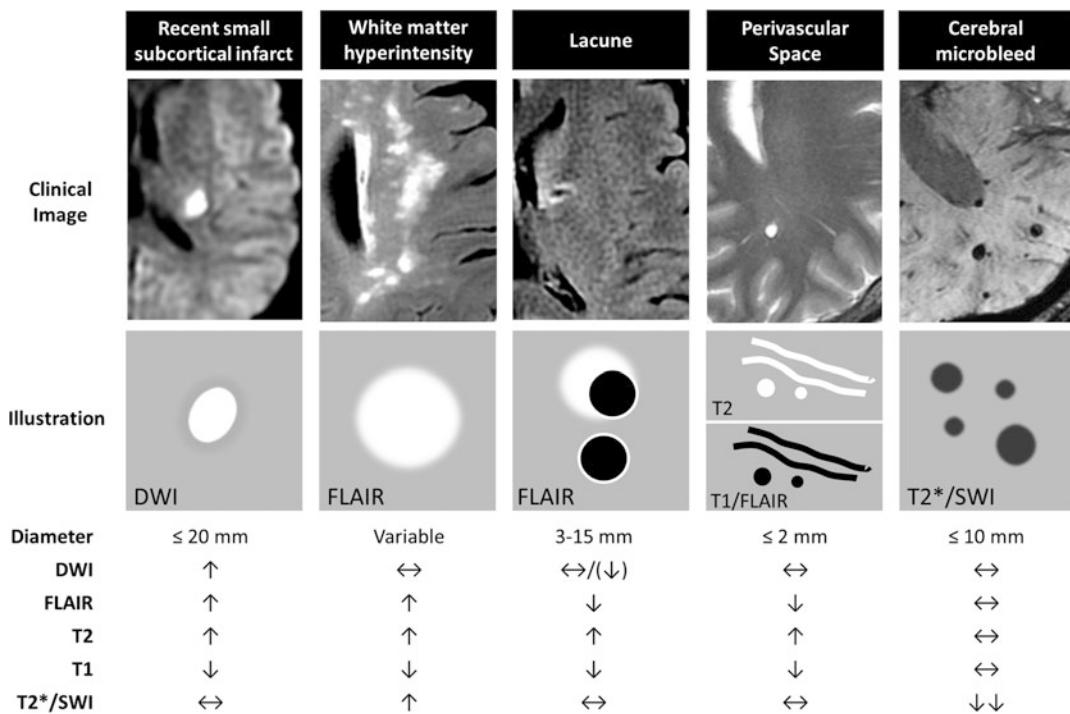


Fig. 4 MR imaging features for cerebral small vessel disease. (Upper) Clinical images (upper) and illustrations (middle) of MRI features for cerebral small vessel disease, with a summary of imaging characteristics (lower) for individual features. DWI, diffusion-weighted imaging. FLAIR, fluid-attenuated inversion recovery. SWI, susceptibility-weighted imaging. ↑, increased signal. ↓, decreased signal. ↔, iso-intense signal. (The figure is reproduced based on reference Wardlaw et al. [145])

quantification of the cSVD imaging features. A “total cSVD score” of the brain could be calculated by combining all pertinent features and may better represent the disease status and burden of cSVD. Such applications could help with disease diagnosis, treatment, monitoring, and prognostication in patients with cSVD.

We will review current machine learning applications for the detection and quantification of cSVD imaging features, including WMH, CMB, lacune, and PVS, as well as the total burden of cSVD.

6.2 White Matter Hyperintensity Segmentation

WMH of presumed vascular origin, characterized by hyperintense lesions on fluid-attenuated inversion recovery (FLAIR) MRI within the white matter, is one of the main features of cSVD [144]. These abnormalities play a key role in normal aging, dementia, and stroke. Large longitudinal population-based studies have confirmed a dose-dependent relationship between WMH volume and clinical outcome, making its measurement of clinical interest [156]. The Fazekas visual rating scale is the most widely used method to assess WMH burden in the clinical setting; it is a four-grade scale rating the size and confluence of WMH lesions in periventricular and deep white matter (Fig. 5) [157]. However, the Fazekas scale has high

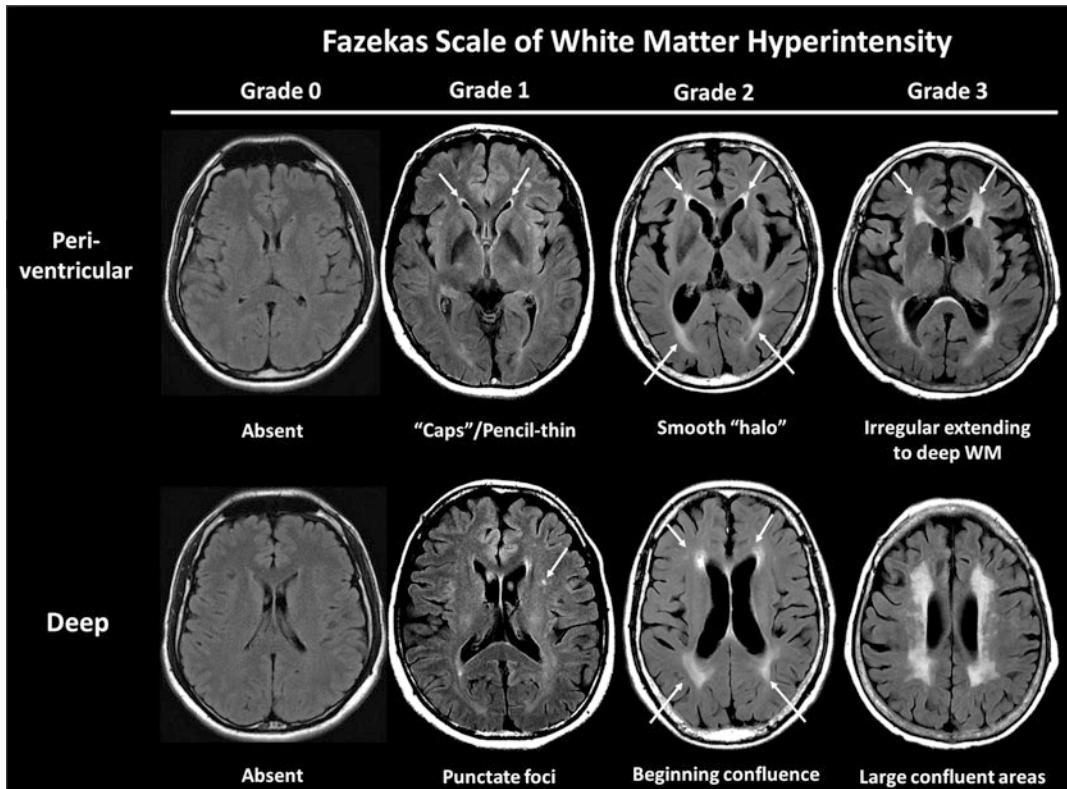


Fig. 5 The Fazekas visual rating scale for white matter hyperintensity. A four-grade scale depending on the size and confluence of lesions is given in the periventricular (upper) and deep white matter (lower) regions, respectively

intra- and inter-subject variability [158], significant ceiling/floor effects [159], and poor sensitivity to clinical group differences [160], leading to inconsistencies in WMH research.

Segmentation and quantification of WMH lesion volume are needed. Before the emergence of deep learning techniques, many automatic WMH segmentation methods were proposed, including supervised methods, e.g., k-nearest neighbors [161], support vector machine [162], Bayesian method based on signal intensity and spatial information [162] or multi-contrast image [163], combined morphological segmentation and adaptive boosting classifier [164], and artificial neural network [165], and unsupervised method, e.g., histogram analysis [166], fuzzy classification algorithm [167, 168], Gaussian mixture model [169], and hidden Markov random field model [170]. However, these methods were generally limited to specific imaging modalities and patient characteristics (e.g., age, clinical presentation) and used different metrics for analysis, making it hard to compare methods to one another [171].

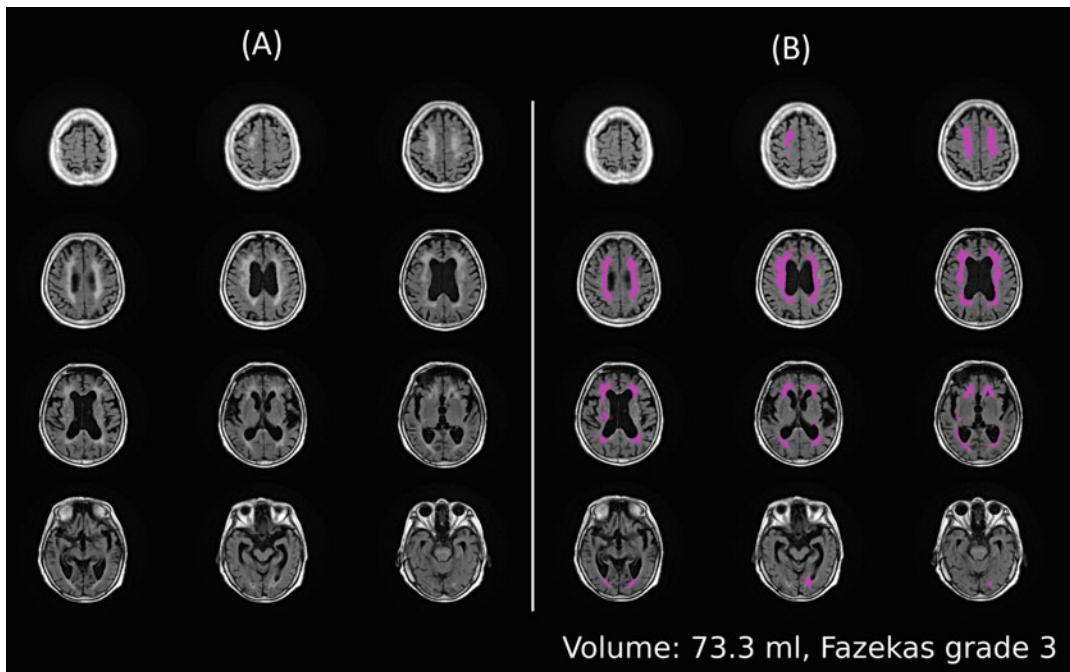


Fig. 6 Example of white matter hyperintensity (WMH) segmentation and quantification. **(a)** The original T2 FLAIR image. **(b)** Automatic WMH segmentation (pink areas) and volume quantification can be achieved by deep learning algorithm which provides a more precise estimation of WMH burden in the brain than the Fazekas scale. WMH white matter hyperintensity

6.2.1 Deep Learning-Based Methods for WMH Segmentation

The WMH Segmentation Challenge at the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) 2017 (<https://wmh.isi.uu.nl/>) provided a standardized assessment of automatic methods for WMH segmentation. The multi-center/multi-scanner dataset comprised images from patients with various degrees of age-related degenerative and vascular pathologies. The training dataset included 60 images from 3 scanners, with manual WMH segmentation by 2 experts as the ground truth. The testing dataset included 110 images obtained from 5 MR scanners, including data from 2 scanners not used in the training set, to evaluate the generalizability of segmentation methods on untested (?) scanners. Five evaluation metrics, including DSC, modified Hausdorff distance, volume difference, sensitivity, and F1 for detecting individual lesion, were used to rank the methods. Among the 20 participants, all the top 10 participants applied deep learning methods [172]. The top-ranking methods performed similarly or better than the two independent human observers, who did not serve as the raters of the ground truth, suggesting the potential of automatic methods to replace human raters (Fig. 6). Li et al. [173], the winner, achieved a DSC of 0.8 and a recall of 0.84 by utilizing an ensemble of three fully convolutional neural networks similar to U-Net with different initializations. Of note, they removed the

WMH prediction in the first and last 1/8 slices, where false-positive prediction frequently occurred, as a post-processing method. Andermatt et al. [174], in second place, utilized a network based on multi-dimensional gated recurrent units (GRU), trained on 3D patches, to achieve a DSC of 0.78 and a recall of 0.83. Ghafoorian et al. [175], in third place, constructed a multi-scale 2D CNN, trained in tenfolds and selecting the three best performing checkpoints on the training data, to achieve a DSC of 0.77, a recall of 0.73, and the highest F1 score of 0.78. Valverde et al. [176], in fourth, constructed a cascade framework of three 3D CNNs, with the first model to identify candidate lesion voxels, the second to reduce false-positive detections, and the third to perform final WMH segmentation. Overall, challenge results indicate that ensemble methods and strategies for false-positive reduction, including selective sampling WMH mimics, removing slices prone to false positives, and adding false-positive reduction model, are advantageous. The top-ranking models generally had very few false positives in normal areas that are hyperintense on FLAIR but are not WMH (e.g., the septum pellucidum), a fault of many lower-ranking methods. Although the top-four ranking models remained to be the leaders in the inter-scanner robustness ranking, some higher-ranking, deep learning-based methods performed worse in inter-scanner robustness than the lower-ranking, rule-based methods, suggesting data-driven approaches sometimes may not generalize well to unseen scanners.

The WMH Segmentation Challenge remains open for new and updated submissions. Zhang et al. [177] designed a dual-path U-Net segmentation model that used an attention mechanism to combine FLAIR sequences and a brain atlas (for location information) inputs to achieve higher performance than the previously mentioned methods. Park et al. [178] proposed a U-Net with multi-scale highlighting foregrounds, which was designed to improve the detection of the WMH voxels with partial volume effects, and achieved a record high of DSC (0.81) and F1 score (0.79).

Although deep learning methods are gaining popularity and have shown great performance in the WMH Segmentation Challenge, a recent systemic review [179] of automatic WMH segmentation methods developed from 2015 to July 2020 showed no evidence to favor deep learning methods in clinical research over the k-NN algorithm [180, 181], linear regression [182, 183], or unsupervised methods (e.g., fuzzy c-means algorithm [184, 185], Gaussian mixture model [186], statistical definition [187]), in terms of spatial agreement with reference segmentations (i.e., DSC). Non-deep learning methods, such as k-NN and linear regression methods, have the advantage of simplicity, can be easier to train, and may be less susceptible to overfitting when dealing

with a limited amount of training data. Future research requires high-quality large-sized open data and code availability to overcome bias in study design and ground truth generation in order to fully compare and validate these methods [188].

6.3 Cerebral Microbleed (CMB) Detection

CMBs are radiological manifestations of cerebral small vessel disease, usually defined as small (≤ 10 mm) areas of signal void on T2*-weighted gradient-recalled echo (GRE) or susceptibility-weighted images (SWI). CMBs are frequently seen in patients with spontaneous intracranial hemorrhage [189] or cognitive impairment [189] and are associated with a higher risk of hemorrhage after IV thrombolysis or therapeutic anticoagulation [190, 191]. CMBs are highly associated with underlying uncontrolled hypertension (particularly when located in deep and/or posterior fossa structures) [192] and/or cerebral amyloid angiopathy (especially when seen in cortical locations) [193]. Detecting CMBs can be clinically important to assess the benefits and risks in treatment planning for stroke patients.

Greenberg et al. [189] published a detailed field guide to CMB detection. The small size of CMBs and the existence of several CMB mimics (e.g., small veins, calcifications, cavernous malformations, iron deposition in deep nucleus, and flow voids) lead to limited inter-observer agreement, long scan interpretation time, and increased error rate by manual inspection, especially for patients with heavy CMB load.

Automatic CMB detection methods might improve the efficiency and accuracy of CMB identification. Radiomic-based and traditional machine learning automatic detection methods have been investigated. Van den Heuvel et al. [194] used morphological features based on the dark and spherical nature of CMBs and random forest classifier to achieve a sensitivity of 89.1% and 25.9 false positives per subject on CMB detection. Several studies have applied deep learning models to improve CMB detection [195–198]. Dou et al. [198] utilized a two-step cascade framework, first with a 3D fully convolutional network for the screening of CMB candidates, followed by a 3D CNN discriminator for the exclusion of CMB mimics, to achieve a sensitivity of 93.16%, precision of 44.31%, and 2.74 false positives per subject for the detection of CMB on SWI. Liu et al. [196] used a two-stage 3D CNN architecture, while adding phase images to SWI as model inputs. The phase images enabled the differentiation of diamagnetic calcifications from paramagnetic CMB, which is not a distinction radiologists can make solely on SWI. Their model successfully reduced false-positive detection and achieved a sensitivity of 95.8%, precision of 70.9%, and 1.6 false positives per subject. Rashid et al. further added quantitative susceptibility mapping (QSM) to SWI as inputs to construct a multi-class U-Net CNN method to differentiate CMBs and non-hemorrhage iron deposits, which was not

achievable with SWI and phase images [197]. The multi-class model reached a sensitivity of 84% and a precision of 59% for CMB detection and a sensitivity of 75% and a precision of 75% for iron deposit detection.

6.4 Lacune Lesion Detection

Lacunes of presumed vascular origin are sequelae of chronic small subcortical infarcts or hemorrhages located in deep gray and white matter in the territory of a perforating arteriole [144]. They are associated with an increased risk of stroke, dementia, and gait impairment [143, 144]. In neuroimaging, lacunes appear as round or ovoid, subcortical, fluid-filled cavities, measuring between 3 and 15 mm, typically showing a surrounding hyperintense gliotic rim on T2 FLAIR images [144]. Longitudinal spatial mapping studies show new WMH forming around small subcortical infarcts [199] and new lacunes forming at the margin of WMH [200], suggesting a strong association and vicinity between the two types of lesions. Therefore, automatic applications that can not only segment WMH but also detect lacunes are desired. However, few studies have proposed automatic methods for lacune detection. Uchiyama et al. [201] developed an algorithm that first used top-hat transformation and multiple-phase binarization techniques to detect potential candidates of lacune and then used rule-based schemes and a support vector machine to eliminate the false positives to achieve a sensitivity of 96.8% with 0.76 false positive per slice. Wang et al. [169] applied a multi-step algorithm to detect WMH, cortical infarcts, and lacunes. The steps included extraction of brain tissue, segmentation of hyperintense lesions from brain tissue using Gaussian mixture model, separation of WMH and cortical infarct based on anatomical location and morphological operation, and segmentation of lacunes based on location and intensity threshold. They achieved a sensitivity of 83.3% with 0.06 false positives per subject for lacune detection. Ghafoorian et al. [202] used a two-stage deep learning method, which included a fully convolutional neural network for candidate detection and a 3D multi-scale location-aware CNN for false-positive reduction. The method achieved a sensitivity of 97.4% with 0.13 false positives per slice.

6.5 Perivascular Space Quantification

Perivascular spaces (PVS), also known as Virchow-Robin spaces, are extensions of extracerebral fluid spaces that surround the penetrating vessels of the brain [144]. They were recently recognized as parts of the glymphatic system, which is a brain-wide perivascular fluid transport system responsible for the clearance of waste in the brain [203]. Normal PVS are not typically seen on conventional MRI, while enlarged PVS are associated with progression of subcortical infarcts, WMH, CMBs, and cognitive decline and are considered a biomarker for cSVD [204]. In neuroimaging, PVS appear as round or ovoid cavities with diameters less than

3 mm and demonstrate signal intensity identical to that of CSF. They are typically located in the inferior basal ganglia, centrum semiovale, and midbrain. PVS may look similar to lacunes on MRI. However, PVS do not have a surrounding gliotic rim and appear more elongated when imaged parallel to the course of the penetrating vessel. The severity of PVS can be graded by a widely used visual rating scale according to Charidimou et al., which is a four-point grade based on the total number of PVS (0, no PVS; 1 [mild], 1–10 PVS; 2 [moderate], 11–20 PVS; 3 [moderate to severe], 21–40 PVS; 4 [severe], > 40 PVS) in the basal ganglia and centrum semiovale [205]. Given the small size and the large number of PVS, it is extremely laborious and time-consuming to perform manual counting or segmentation of PVS, which may explain the scarcity of studies about automatic methods for PVS quantification in the literature. Park et al. [206] proposed a supervised method to perform automatic PVS segmentation method based on manually derived PVS masks on 7 T MR images. They extracted Haar-like features, which are often used in object recognition, from regions of interest determined by brain and vascular structure and used a random forest classifier to achieve a DSC of 0.73, sensitivity of 69%, and positive predictive value of 80%. Ballerini et al. [207] propose a PVS segmentation technique based on the 3D Frangi filtering. Because of the lack of ground truth of PVS segmentation mask, they alternatively optimized and evaluated the method by using ordered logit models and visual rating scales. The method achieved a Spearman's correlation coefficient of 0.74 ($p < 0.001$) between segmentation-based PVS burden and visual rating scale. Dubost et al. [208] used 3D convolutional neural network regression to predict visual rating scale and achieved an intraclass correlation coefficient of 0.75–0.88 between visual and automated scales, which was even higher than the inter-observer agreement among human raters.

6.6 Total Small Vessel Disease Burden

cSVD is considered a dynamic, whole brain disorder with a wide spectrum of clinical presentations and diffuse imaging manifestations in the brain while sharing common microvascular pathologies [209]. A multifactorial approach that combines all imaging features may better represent the burden and disease status of cSVD. Several visual scoring systems of total cSVD burden have been introduced [154, 205]. Staals et al. [154] proposed a four-point score in which one point is given in the presence of each of the cSVD imaging feature: (1) more than one lacune, (2) more than one microbleed, (3) moderate to severe (more than 11) PVS in basal ganglia, and (4) periventricular WMH Fazekas score of 3 and/or deep WMH Fazekas score of 2–3. Although these semiquantitative scoring systems are pragmatic and simple for clinical use, they have several limitations. First, they may not be sensitive enough to represent the severity of the disease, as the accumulation of cSVD burden forms a

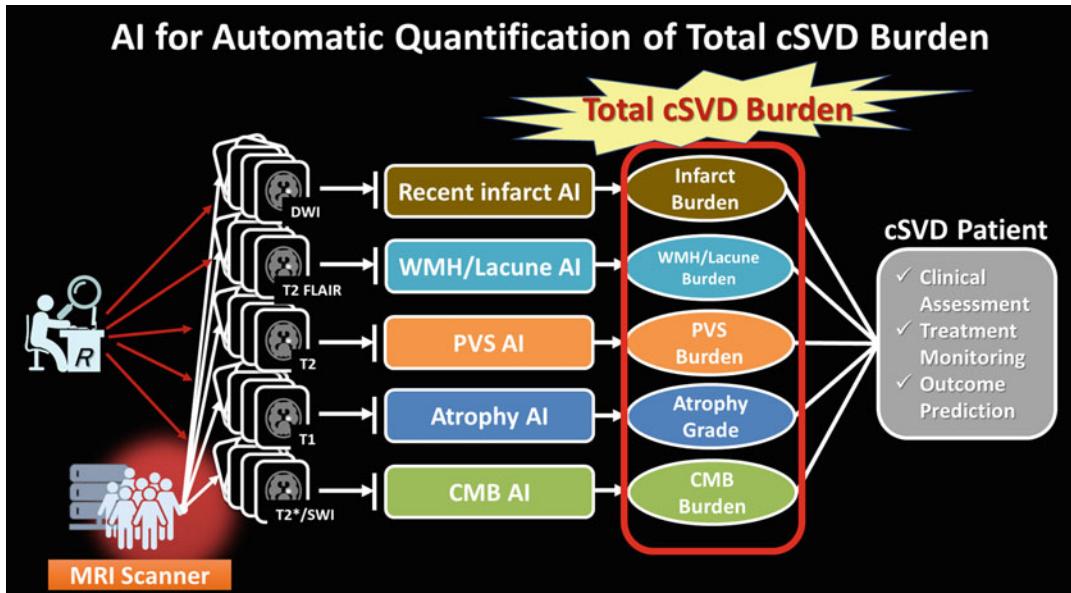


Fig. 7 AI applications for cerebral small vessel disease (cSVD). AI algorithms have great potential to perform automatic quantification of individual cSVD imaging features. By combining these burdens, a “total cSVD burden” could be quantified, which might facilitate the clinical assessment, treatment monitoring, and outcome prediction in patients with cSVD

continuum, rather than several ordinal scores. Second, visual scoring may be subjective and laborious for raters, especially for WMH and PVS evaluation. Third, existing scoring doesn’t account for lesion location, but anatomical location is a known key factor for cognitive impairment [210]. The automatic methods for different cSVD imaging features described in the previous sections can offer quantitative measurements of the cSVD burden in the whole brain and are well suited to overcome these limitations. Several studies have shown great potential for computer-generated total cSVD burden in the assessment of cSVD patients. Duan et al. [211] developed a multiple CNN-based system that can accurately segment subcortical infarcts, CMBs, WMHs, and lacunes 4.4 s per subject. Dickie et al. [212] used a voxel-based Gaussian mixture model cluster analysis on multi-contrast MR images to estimate overall WMH, lacunes, CMBs, and atrophy into a “brain health index”; they showed the brain health index has a stronger association with cognitive outcome than WMH volume and visual cSVD score. Jokinen et al. [213] used automated atlas- and CNN-based segmentation methods to yield volumetric measures of WMHs, lacunes, PVS, cortical infarcts, and brain atrophy to show that the combined measure of all markers was a more powerful predictor of cognitive and functional outcomes than any individual measure alone.

Overall, previous studies have shown great potential of machine learning algorithms to perform automatic segmentation or detection of cSVD imaging features. By combining the measurement of each cSVD feature, a “total cSVD burden” can be quantified, which might be used to facilitate clinical assessment, treatment monitoring, and outcome prediction in patients with cSVD (Fig. 7).

7 Conclusion

In conclusion, machine learning algorithms show great potential in improving clinical diagnosis and care for cerebrovascular disorders. ML performance varies by study and dataset, but in many cases already exceeds the current clinical state-of-the-art [?measures]. There is a need for more large cohort validation studies, and the development of standard test sets for comparing different algorithms would enable fairer comparison between methods. In addition, more real-world experience is necessary to understand the role of machine learning in improving the diagnosis and care of cerebrovascular disorders.

Acknowledgments

The authors are grateful to Margy McCullough-Hicks for her insightful comments on the chapter.

References

1. Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW et al (2021) Heart disease and stroke Statistics-2021 update: a report from the American Heart Association. *Circulation* 143(8):e254–e743. <https://doi.org/10.1161/CIR.0000000000000950>
2. Collaborators GBDS (2021) Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the global burden of disease study 2019. *Lancet Neurol* 20(10):795–820. [https://doi.org/10.1016/S1474-4422\(21\)00252-0](https://doi.org/10.1016/S1474-4422(21)00252-0)
3. Cannistraro RJ, Badi M, Eidelberg BH, Dickson DW, Middlebrooks EH, Meschia JF (2019) CNS small vessel disease: a clinical review. *Neurology* 92(24):1146–1156. <https://doi.org/10.1212/WNL.0000000000007654>
4. Ingall T, Asplund K, Mähönen M, Bonita R (2000) A multinational comparison of subarachnoid hemorrhage epidemiology in the WHO MONICA stroke study. *Stroke* 31(5):1054–1061. <https://doi.org/10.1161/01.str.31.5.1054>
5. van Gijn J, Kerr RS, Rinkel GJE (2007) Subarachnoid haemorrhage. *Lancet* (London, England) 369(9558):306–318. [https://doi.org/10.1016/S0140-6736\(07\)60153-6](https://doi.org/10.1016/S0140-6736(07)60153-6)
6. Hop JW, Rinkel GJ, Algra A, van Gijn J (1997) Case-fatality rates and functional outcome after subarachnoid hemorrhage: a systematic review. *Stroke* 28(3):660–664. <https://doi.org/10.1161/01.str.28.3.660>
7. Saver JL, Goyal M, van der Lugt A, Menon BK, Majoe CB, Dippel DW et al (2016) Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: a meta-analysis. *JAMA* 316(12):1279–1288. <https://doi.org/10.1001/jama.2016.13647>
8. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K et al (2019) Guidelines for the early management of patients with acute ischemic stroke: 2019

- update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 50(12):e344–e418. <https://doi.org/10.1161/STR.000000000000211>
9. Rennert RC, Wali AR, Steinberg JA, Santiago-Dieppa DR, Olson SE, Pannell JS et al (2019) Epidemiology, natural history, and clinical presentation of large vessel ischemic stroke. *Neurosurgery* 85(suppl_1):S4–S8. <https://doi.org/10.1093/neuros/nyz042>
 10. Goyal M, Menon BK, van Zwam WH, Dippel DW, Mitchell PJ, Demchuk AM et al (2016) Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. *Lancet* 387(10029):1723–1731. [https://doi.org/10.1016/S0140-6736\(16\)00163-X](https://doi.org/10.1016/S0140-6736(16)00163-X)
 11. Badhiwala JH, Nassiri F, Alhazzani W, Selim MH, Farrokhyar F, Spears J et al (2015) Endovascular thrombectomy for acute ischemic stroke: a meta-analysis. *JAMA* 314(17): 1832–1843. <https://doi.org/10.1001/jama.2015.13767>
 12. Hassan AE, Ringheanu VM, Rabah RR, Preston L, Tekle WG, Qureshi AI (2020) Early experience utilizing artificial intelligence shows significant reduction in transfer times and length of stay in a hub and spoke model. *Interv Neuroradiol* 26(5):615–622. <https://doi.org/10.1177/1591019920953055>
 13. Yahav-Dovrat A, Saban M, Merhav G, Lankri I, Abergel E, Eran A et al (2021) Evaluation of artificial intelligence-powered identification of large-vessel occlusions in a comprehensive stroke center. *AJNR Am J Neuroradiol* 42(2):247–254. <https://doi.org/10.3174/ajnr.A6923>
 14. Stib MT, Vasquez J, Dong MP, Kim YH, Subzwari SS, Triedman HJ et al (2020) Detecting large vessel occlusion at multiphase CT angiography by using a deep convolutional neural network. *Radiology* 297(3): 640–649. <https://doi.org/10.1148/radiol.2020200334>
 15. Dehkharghani S, Lansberg M, Venkatsubramanian C, Cereda C, Lima F, Coelho H et al (2021) High-performance automated anterior circulation CT angiographic clot detection in acute stroke: a multi-reader comparison. *Radiology* 298(3): 665–670. <https://doi.org/10.1148/radiol.2021202734>
 16. You J, Yu PLH, Tsang ACO, Tsui ELH, Woo PPS, Leung GKK (2019) Automated computer evaluation of acute ischemic stroke and large vessel occlusion. arXiv preprint arXiv:1906.08059
 17. You J, Tsang ACO, Yu PLH, Tsui ELH, Woo PPS, Lui CSM et al (2020) Automated hierarchy evaluation system of large vessel occlusion in acute ischemia stroke. *Front Neuroinform* 14:13. <https://doi.org/10.3389/fninf.2020.00013>
 18. Olive-Gadea M, Crespo C, Granes C, Hernandez-Perez M, Perez de la Ossa N, Laredo C et al (2020) Deep learning based software to identify large vessel occlusion on noncontrast computed tomography. *Stroke* 51(10):3133–3137. <https://doi.org/10.1161/STROKEAHA.120.030326>
 19. Gaha M, Roy C, Estrade L, Gevry G, Weill A, Roy D et al (2014) Inter- and intraobserver agreement in scoring angiographic results of intra-arterial stroke therapy. *AJNR Am J Neuroradiol* 35(6):1163–1169. <https://doi.org/10.3174/ajnr.A3828>
 20. Volny O, Cimflova P, Szeder V (2017) Inter-Rater reliability for thrombolysis in cerebral infarction with TICI 2c category. *J Stroke Cerebrovasc Dis* 26(5):992–994. <https://doi.org/10.1016/j.jstrokecerebrovasdis.2016.11.008>
 21. Ueda D, Katayama Y, Yamamoto A, Ichinose T, Arima H, Watanabe Y et al (2021) Deep learning-based angiogram generation model for cerebral angiography without misregistration artifacts. *Radiology* 299(3):675–681. <https://doi.org/10.1148/radiol.2021203692>
 22. Zhang M, Zhang C, Wu X, Cao X, Young GS, Chen H et al (2020) A neural network approach to segment brain blood vessels in digital subtraction angiography. *Comput Methods Prog Biomed* 185:105159. <https://doi.org/10.1016/j.cmpb.2019.105159>
 23. Bhurwani MMS, Snyder KV, Waqas M, Mokin M, Rava RA, Podgorsak AR et al (2021) Use of biplane quantitative angiographic imaging with ensemble neural networks to assess reperfusion status during mechanical thrombectomy. *Proc SPIE Int Soc Opt Eng* 11597:115971F
 24. Su R, Cornelissen SAP, van der Sluijs M, van Es A, van Zwam WH, Dippel DWJ et al (2021) autoTICI: automatic brain tissue reperfusion scoring on 2D DSA images of acute ischemic stroke patients. *IEEE Trans Med Imaging* 40(9):2380–2391. <https://doi.org/10.1109/TMI.2021.307713>

25. Su R, van der Sluijs M, Cornelissen SAP, Lycklama G, Hofmeijer J, Majoe C et al (2022) Spatio-temporal deep learning for automatic detection of intracranial vessel perforation in digital subtraction angiography during endovascular thrombectomy. *Med Image Anal* 77:102377. <https://doi.org/10.1016/j.media.2022.102377>
26. Wang J, Zhang J, Gong X, Zhang W, Zhou Y, Lou M (2022) Prediction of large vessel occlusion for ischaemic stroke by using the machine learning model random forests. *Stroke Vasc Neurol* 7(2):94–100. <https://doi.org/10.1136/svn-2021-001096>
27. Chen Z, Zhang R, Xu F, Gong X, Shi F, Zhang M et al (2018) Novel prehospital prediction model of large vessel occlusion using artificial neural network. *Front Aging Neurosci* 10:181. <https://doi.org/10.3389/fnagi.2018.00181>
28. Tarkanyi G, Tenyi A, Hollos R, Kalmar PJ, Szapary L (2022) Optimization of large vessel occlusion detection in acute ischemic stroke using machine learning methods. *Life (Basel)* 12(2):230. <https://doi.org/10.3390/life12020230>
29. Uchida K, Kouno J, Yoshimura S, Kinjo N, Sakakibara F, Araki H et al (2022) Development of machine learning models to predict probabilities and types of stroke at prehospital stage: the Japan urgent stroke triage score using machine learning (JUST-ML). *Transl Stroke Res* 13(3):370–381. <https://doi.org/10.1007/s12975-021-00937-x>
30. Hayashi Y, Shimada T, Hattori N, Shimazui T, Yoshida Y, Miura RE et al (2021) A prehospital diagnostic algorithm for strokes using machine learning: a prospective observational study. *Sci Rep* 11(1): 20519. <https://doi.org/10.1038/s41598-021-99828-2>
31. Thomalla G, Simonsen CZ, Boutitie F, Andersen G, Berthezene Y, Cheng B et al (2018) MRI-guided thrombolysis for stroke with unknown time of onset. *N Engl J Med* 379(7):611–622. <https://doi.org/10.1056/NEJMoa1804355>
32. Mackey J, Kleindorfer D, Sucharew H, Moonaw CJ, Kissela BM, Alwell K et al (2011) Population-based study of wake-up strokes. *Neurology* 76(19):1662–1667. <https://doi.org/10.1212/WNL.0b013e318219fb30>
33. Fink JN, Kumar S, Horkan C, Linfante I, Selim MH, Caplan LR et al (2002) The stroke patient who woke up: clinical and radiological features, including diffusion and perfusion MRI. *Stroke* 33(4):988–993. <https://doi.org/10.1161/01.str.0000014585.17714.67>
34. Xu XQ, Zu QQ, Lu SS, Cheng QG, Yu J, Sheng Y et al (2014) Use of FLAIR imaging to identify onset time of cerebral ischemia in a canine model. *AJNR Am J Neuroradiol* 35(2):311–316. <https://doi.org/10.3174/ajnr.A3689>
35. Thomalla G, Rossbach P, Rosenkranz M, Siemonsen S, Krutzellmann A, Fiehler J et al (2009) Negative fluid-attenuated inversion recovery imaging identifies acute ischemic stroke at 3 hours or less. *Ann Neurol* 65(6): 724–732. <https://doi.org/10.1002/ana.21651>
36. Petkova M, Rodrigo S, Lamy C, Openheim G, Touze E, Mas JL et al (2010) MR imaging helps predict time from symptom onset in patients with acute stroke: implications for patients with unknown onset time. *Radiology* 257(3):782–792. <https://doi.org/10.1148/radiol.10100461>
37. Lee H, Lee EJ, Ham S, Lee HB, Lee JS, Kwon SU et al (2020) Machine learning approach to identify stroke within 4.5 hours. *Stroke* 51(3): 860–866. <https://doi.org/10.1161/STROKEAHA.119.027611>
38. Zhu H, Jiang L, Zhang H, Luo L, Chen Y, Chen Y (2021) An automatic machine learning approach for ischemic stroke onset time identification based on DWI and FLAIR imaging. *Neuroimage Clin* 31: 102744. <https://doi.org/10.1016/j.nicl.2021.102744>
39. Ho KC, Speier W, Zhang H, Scalzo F, El-Saden S, Arnold CW (2019) A machine learning approach for classifying ischemic stroke onset time from imaging. *IEEE Trans Med Imaging* 38(7):1666–1676. <https://doi.org/10.1109/TMI.2019.2901445>
40. Ho KC, Speier W, El-Saden S, Arnold CW (2017) Classifying acute ischemic stroke onset time using deep imaging features. *AMIA Annu Symp Proc* 2017:892–901
41. Pexman JH, Barber PA, Hill MD, Sevick RJ, Demchuk AM, Hudon ME et al (2001) Use of the Alberta stroke program early CT score (ASPECTS) for assessing CT scans in patients with acute stroke. *AJNR Am J Neuroradiol* 22(8):1534–1542
42. Yoshimura S, Sakai N, Yamagami H, Uchida K, Beppu M, Toyoda K et al (2022) Endovascular therapy for acute stroke with a large ischemic region. *N Engl J Med*. <https://doi.org/10.1056/NEJMoa2118191>
43. Chen L, Bentley P, Rueckert D (2017) Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. *Neuroimage Clin* 15:633–643. <https://doi.org/10.1016/j.nicl.2017.06.016>

44. Wu O, Winzeck S, Giese AK, Hancock BL, Etherton MR, Bouts M et al (2019) Big data approaches to phenotyping acute ischemic stroke using automated lesion segmentation of multi-Center magnetic resonance imaging data. *Stroke* 50(7):1734–1741. <https://doi.org/10.1161/STROKEAHA.119.025373>
45. Maier O, Menze BH, von der Gabeltz J, Hani L, Heinrich MP, Liebrand M et al (2017) ISLES 2015 - a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Med Image Anal* 35:250–269. <https://doi.org/10.1016/j.media.2016.07.009>
46. Praveen GB, Agrawal A, Sundaram P, Sardesai S (2018) Ischemic stroke lesion segmentation using stacked sparse autoencoder. *Comput Biol Med* 99:38–52. <https://doi.org/10.1016/j.combiomed.2018.05.027>
47. Tomita N, Jiang S, Maeder ME, Hassanpour S (2020) Automatic post-stroke lesion segmentation on MR images using 3D residual convolutional neural network, vol 27, p 102276
48. Liu L, Kurgan L, Wu FX, Wang J (2020) Attention convolutional neural network for accurate segmentation and quantification of lesions in ischemic stroke disease. *Med Image Anal* 65:101791. <https://doi.org/10.1016/j.media.2020.101791>
49. Karthik R, Gupta U, Jha A, Rajalakshmi R, Menaka R (2019) A deep supervised approach for ischemic lesion segmentation from multimodal MRI using fully convolutional network. *Appl Soft Comput* 84:105685. <https://doi.org/10.1016/j.asoc.2019.105685>
50. Liu L, Chen S, Zhang F, Wu F, Pan Y, Wang J (2020) Deep convolutional neural network for automatically segmenting acute ischemic stroke lesion in multi-modality MRI. *Neural Comput Appl* 32:6545
51. Zhang L, Song R, Wang Y, Zhu C, Liu J, Yang J et al (2020) Ischemic stroke lesion segmentation using multi-plane information fusion. *IEEE Access* 8:45715–45725. <https://doi.org/10.1109/ACCESS.2020.2977415>
52. Zhao B, Ding S, Wu H, Liu G, Cao C, Jin S, et al. (2019) Automatic acute ischemic stroke lesion segmentation using semi-supervised learning. <https://doi.org/10.48550/arXiv.1908.03735>
53. Federau C, Christensen S, Scherrer N, Ospel JM, Schulze-Zachau V, Schmidt N et al (2020) Improved segmentation and detection sensitivity of diffusion-weighted stroke lesions with synthetically enhanced deep learning. *Radiol Artif Intell* 2(5):e190217. <https://doi.org/10.1148/ryai.2020190217>
54. Woo I, Lee A, Jung SC, Lee H, Kim N, Cho SJ et al (2019) Fully automatic segmentation of acute ischemic lesions on diffusion-weighted imaging using convolutional neural networks: comparison with conventional algorithms. *Korean J Radiol* 20(8):1275–1284. <https://doi.org/10.3348/kjr.2018.0615>
55. Winzeck S, Mocking SJT, Bezerra R, Bouts M, McIntosh EC, Diwan I et al (2019) Ensemble of convolutional neural networks improves automated segmentation of acute ischemic lesions using multiparametric diffusion-weighted MRI. *AJNR Am J Neuroradiol* 40(6):938–945. <https://doi.org/10.3174/ajnr.A6077>
56. Do LN, Baek BH, Kim SK, Yang HJ, Park I, Yoon W (2020) Automatic assessment of ASPECTS using diffusion-weighted imaging in acute ischemic stroke using recurrent residual convolutional neural network. *Diagnostics (Basel)* 10(10):803. <https://doi.org/10.3390/diagnostics10100803>
57. Qiu W, Kuang H, Teleg E, Ospel JM, Sohn SI, Almekhlafi M et al (2020) Machine learning for detecting early infarction in acute stroke with non-contrast-enhanced CT. *Radiology* 294(3):638–644. <https://doi.org/10.1148/radiol.2020191193>
58. Kuang H, Najm M, Chakraborty D, Maraj N, Sohn SI, Goyal M et al (2019) Automated ASPECTS on noncontrast CT scans in patients with acute ischemic stroke using machine learning. *AJNR Am J Neuroradiol* 40(1):33–38. <https://doi.org/10.3174/ajnr.A5889>
59. Albers GW, Wald MJ, Mlynash M, Endres J, Bammer R, Straka M et al (2019) Automated calculation of Alberta stroke program early CT score: validation in patients with large hemispheric infarct. *Stroke* 50(11):3277–3279. <https://doi.org/10.1161/STROKEAHA.119.026430>
60. Maegerlein C, Fischer J, Monch S, Berndt M, Wunderlich S, Seifert CL et al (2019) Automated calculation of the Alberta stroke program early CT score: feasibility and reliability. *Radiology* 291(1):141–148. <https://doi.org/10.1148/radiol.2019181228>
61. Brinjikji W, Abbasi M, Arnold C, Benson JC, Braksick SA, Campeau N et al (2021) e-ASPECTS software improves interobserver agreement and accuracy of interpretation of aspects score. *Interv Neuroradiol* 27(6):781–787. <https://doi.org/10.1177/15910199211011861>
62. Neuhaus A, Seyedsaadat SM, Mihal D, Benson J, Mark I, Kallmes DF et al (2020)

- Region-specific agreement in ASPECTS estimation between neuroradiologists and e-ASPECTS software. *J Neurointerv Surg.* 12(7):720–723. <https://doi.org/10.1136/neurintsurg-2019-015442>
63. Nagel S, Sinha D, Day D, Reith W, Chapot R, Papanagiotou P et al (2017) e-ASPECTS software is non-inferior to neuroradiologists in applying the ASPECT score to computed tomography scans of acute ischemic stroke patients. *Int J Stroke* 12(6):615–622. <https://doi.org/10.1177/1747493016681020>
64. Herweh C, Ringleb PA, Rauch G, Gerry S, Behrens L, Mohlenbruch M et al (2016) Performance of e-ASPECTS software in comparison to that of stroke physicians on assessing CT scans of acute ischemic stroke patients. *Int J Stroke* 11(4):438–445. <https://doi.org/10.1177/1747493016632244>
65. Hakim A, Christensen S, Winzeck S, Lansberg MG, Parsons MW, Lucas C et al (2021) Predicting infarct core from computed tomography perfusion in acute ischemia with machine learning: lessons from the ISLES challenge. *Stroke* 52(7):2328–2337. <https://doi.org/10.1161/STROKEAHA.120.030696>
66. Song T (2019) Generative model-based ischemic stroke lesion segmentation. <https://doi.org/10.48550/arXiv.1906.02392>
67. Clerigues A, Valverde S, Bernal J, Freixenet J, Oliver A, Llado X (2019) Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks. *Comput Biol Med* 115:103487. <https://doi.org/10.1016/j.combiomed.2019.103487>
68. Mazdak Abulnaga S, Rubin J (2018) Ischemic stroke lesion segmentation in CT perfusion scans using pyramid pooling and focal loss. <https://doi.org/10.48550/arXiv.1811.01085>
69. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2016) Feature pyramid networks for object detection. <https://doi.org/10.48550/arXiv.1612.03144>
70. Winzeck S, Hakim A, McKinley R, Pinto J, Alves V, Silva C et al (2018) ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. *Front Neurol* 9:679. <https://doi.org/10.3389/fneur.2018.00679>
71. Pinto A, McKinley R, Alves V, Wiest R, Silva CA, Reyes M (2018) Stroke lesion outcome prediction based on MRI imaging combined with clinical information. *Front Neurol* 9:1060. <https://doi.org/10.3389/fneur.2018.01060>
72. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014
73. Nielsen A, Hansen MB, Tietze A, Mouridsen K (2018) Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning. *Stroke* 49(6):1394–1401. <https://doi.org/10.1161/STROKEAHA.117.019740>
74. Ho KC, Scalzo F, Sarma KV, Speier W, El-Saden S, Arnold C (2019) Predicting ischemic stroke tissue fate using a deep convolutional neural network on source magnetic resonance perfusion images. *J Med Imaging (Bellingham)* 6(2):026001. <https://doi.org/10.1117/1.JMI.6.2.026001>
75. Yu Y, Xie Y, Thamm T, Gong E, Ouyang J, Huang C et al (2020) Use of deep learning to predict final ischemic stroke lesions from initial magnetic resonance imaging. *JAMA Netw Open* 3(3):e200772. <https://doi.org/10.1001/jamanetworkopen.2020.0772>
76. Yu Y, Xie Y, Thamm T, Gong E, Ouyang J, Christensen S et al (2021) Tissue at risk and ischemic core estimation using deep learning in acute stroke. *AJNR Am J Neuroradiol*. <https://doi.org/10.3174/ajnr.A7081>
77. Wang K, Shou Q, Ma SJ, Liebeskind D, Qiao XJ, Saver J et al (2020) Deep learning detection of penumbral tissue on arterial spin labeling in stroke. *Stroke* 51(2):489–497. <https://doi.org/10.1161/STROKEAHA.119.027457>
78. Robben D, Boers AMM, Marquering HA, Langezaal L, Roos Y, van Oostenbrugge RJ et al (2020) Prediction of final infarct volume from native CT perfusion and treatment parameters using deep learning. *Med Image Anal* 59:101589. <https://doi.org/10.1016/j.media.2019.101589>
79. Amador K, Wilms M, Winder A, Fiehler J, Forkert N (2021) Stroke lesion outcome prediction based on 4D CT perfusion data using temporal convolutional networks. In: Mattias H, Qi D, Marleen de B, Jan L, Alexander S, Floris E (eds) Proceedings of the fourth conference on medical imaging with deep learning. Proceedings of Machine Learning Research. PMLR, pp 22–33
80. Kuang H, Qiu W, Boers AM, Brown S, Muir K, Majoe C et al (2021) Computed tomography perfusion-based machine learning model better predicts follow-up infarction in patients with acute ischemic stroke. *Stroke* 52(1):223–231. <https://doi.org/10.1161/STROKEAHA.120.030092>

81. Yu Y, Parsi B, Speier W, Arnold C, Lou M, Scalzo F (2019) LSTM network for prediction of hemorrhagic transformation in acute stroke. Medical image computing and computer-assisted intervention. Springer, Cham, pp 177–185
82. Yu Y, Guo D, Lou M, Liebeskind D, Scalzo F (2018) Prediction of hemorrhagic transformation severity in acute stroke from source perfusion MRI. IEEE Trans Biomed Eng 65(9):2058–2065. <https://doi.org/10.1109/TBME.2017.2783241>
83. Jiang L, Zhou L, Yong W, Cui J, Geng W, Chen H et al (2021) A deep learning-based model for prediction of hemorrhagic transformation after stroke. Brain Pathol:e13023. <https://doi.org/10.1111/bpa.13023>
84. Nishi H, Oishi N, Ishii A, Ono I, Ogura T, Sunohara T et al (2019) Predicting clinical outcomes of large vessel occlusion before mechanical thrombectomy using machine learning. Stroke 50(9):2379–2388. <https://doi.org/10.1161/STROKEAHA.119.025411>
85. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH (2019) Machine learning-based model for prediction of outcomes in acute stroke. Stroke 50(5):1263–1265. <https://doi.org/10.1161/STROKEAHA.118.024293>
86. Ntaios G, Faouzi M, Ferrari J, Lang W, Vemmos K, Michel P (2012) An integer-based score to predict functional outcome in acute ischemic stroke: the ASTRAL score. Neurology 78(24):1916–1922. <https://doi.org/10.1212/WNL.0b013e318259e221>
87. Ho KC, Speier W, El-Saden S, Liebeskind DS, Saver JL, Bui AA et al (2014) Predicting discharge mortality after acute ischemic stroke using balanced data. AMIA Annu Symp Proc 2014:1787–1796
88. van Os HJA, Ramos LA, Hilbert A, van Leeuwen M, van Walderveen MAA, Kruijt ND et al (2018) Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. Front Neurol 9:784. <https://doi.org/10.3389/fneur.2018.00784>
89. Xie Y, Jiang B, Gong E, Li Y, Zhu G, Michel P et al (2019) JOURNAL CLUB: use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information. AJR Am J Roentgenol 212(1):44–51. <https://doi.org/10.2214/AJR.18.20260>
90. Osama S, Zafar K, Sadiq MU (2020) Predicting clinical outcome in acute ischemic stroke using parallel multi-parametric feature embedded Siamese network. Diagnostics (Basel) 10(11):858. <https://doi.org/10.3390/diagnostics10110858>
91. Chicco D (2021) Siamese neural networks: an overview. Methods Mol Biol 2190:73–94. https://doi.org/10.1007/978-1-0716-0826-5_3
92. Nishi H, Oishi N, Ishii A, Ono I, Ogura T, Sunohara T et al (2020) Deep learning-derived high-level neuroimaging features predict clinical outcomes for large vessel occlusion. Stroke 51(5):1484–1492. <https://doi.org/10.1161/STROKEAHA.119.028101>
93. Guo J, Gong E, Fan AP, Goubran M, Khalighi MM, Zaharchuk G (2020) Predicting (15)O-Water PET cerebral blood flow maps from multi-contrast MRI using a deep convolutional neural network with evaluation of training cohort bias. J Cereb Blood Flow Metab 40(11):2240–2253. <https://doi.org/10.1177/0271678X19888123>
94. Gupta A, Chazen JL, Hartman M, Delgado D, Anumula N, Shao H et al (2012) Cerebrovascular reserve and stroke risk in patients with carotid stenosis or occlusion: a systematic review and meta-analysis. Stroke 43(11):2884–2891. <https://doi.org/10.1161/STROKEAHA.112.663716>
95. Chen DYT, Ishii Y, Fan AP, Guo J, Zhao MY, Steinberg GK et al (2020) Predicting PET cerebrovascular reserve with deep learning by using baseline MRI: a pilot investigation of a drug-free brain stress test. Radiology 296(3):627–637. <https://doi.org/10.1148/radiol.2020192793>
96. Collaborators GBDS (2019) Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the global burden of disease study 2016. Lancet Neurol 18(5):439–458. [https://doi.org/10.1016/S1474-4422\(19\)30034-1](https://doi.org/10.1016/S1474-4422(19)30034-1)
97. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK et al (2018) Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 392(10162):2388–2396. [https://doi.org/10.1016/S0140-6736\(18\)31645-3](https://doi.org/10.1016/S0140-6736(18)31645-3)
98. Lee H, Yune S, Mansouri M, Kim M, Tajmir SH, Guerrier CE et al (2019) An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nat Biomed Eng 3(3):173–182. <https://doi.org/10.1038/s41551-018-0324-9>
99. Kuo W, Hne C, Mukherjee P, Malik J, Yuh EL (2019) Expert-level detection of acute

- intracranial hemorrhage on head computed tomography using deep learning. *Proc Natl Acad Sci U S A* 116(45):22737–22745. <https://doi.org/10.1073/pnas.1908021116>
100. Zhang Y, Zhang B, Liang F, Liang S, Zhang Y, Yan P et al (2019) Radiomics features on non-contrast-enhanced CT scan can precisely classify AVM-related hematomas from other spontaneous intraparenchymal hematoma types. *Eur Radiol* 29(4):2157–2165. <https://doi.org/10.1007/s00330-018-5747-x>
101. Wang T, Lei Y, Tian S, Jiang X, Zhou J, Liu T et al (2019) Learning-based automatic segmentation of arteriovenous malformations on contrast CT images in brain stereotactic radiosurgery. *Med Phys* 46(7):3133–3141. <https://doi.org/10.1002/mp.13560>
102. Lee CC, Yang HC, Lin CJ, Chen CJ, Wu HM, Shiau CY et al (2019) Intervening nodal brain parenchyma and risk of radiation-induced changes after radiosurgery for brain arteriovenous malformation: a study using an unsupervised machine learning algorithm. *World Neurosurg* 125:e132–e1e8. <https://doi.org/10.1016/j.wneu.2018.12.220>
103. Turan N, Heider RA, Roy AK, Miller BA, Mullins ME, Barrow DL et al (2018) Current perspectives in imaging modalities for the assessment of unruptured intracranial aneurysms: a comparative analysis and review. *World Neurosurg* 113:280–292. <https://doi.org/10.1016/j.wneu.2018.01.054>
104. Yoon NK, McNally S, Taussky P, Park MS (2016) Imaging of cerebral aneurysms: a clinical perspective. *Neurovasc Imaging* 2(1):6. <https://doi.org/10.1186/s40809-016-0016-3>
105. Jaja BNR, Cusimano MD, Etminan N, Hanggi D, Hasan D, Ilodigwe D et al (2013) Clinical prediction models for aneurysmal subarachnoid hemorrhage: a systematic review. *Neurocrit Care* 18(1):143–153. <https://doi.org/10.1007/s12028-012-9792-z>
106. Hemphill JC, Greenberg SM, Anderson CS, Becker K, Bendok BR, Cushman M et al (2015) Guidelines for the management of spontaneous intracerebral hemorrhage: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 46(7):2032–2060. <https://doi.org/10.1161/STR.0000000000000069>
107. Ois A, Vivas E, Figueras-Aguirre G, Guimaraens L, Cuadrado-Godia E, Avellaneda C et al (2019) Misdiagnosis worsens prognosis in subarachnoid hemorrhage with good Hunt and Hess score. *Stroke* 50(11):3072–3076. <https://doi.org/10.1161/STROKEAHA.119.025520>
108. Lubicz B, Levivier M, François O, Thoma P, Sadeghi N, Collignon L et al (2007) Sixty-four-row multisector CT angiography for detection and evaluation of ruptured intracranial aneurysms: interobserver and intertechnique reproducibility. *AJNR Am J Neuroradiol* 28(10):1949–1955. <https://doi.org/10.3174/ajnr.A0699>
109. Okahara M, Kiyosue H, Yamashita M, Nagatomi H, Hata H, Saginoya T et al (2002) Diagnostic accuracy of magnetic resonance angiography for cerebral aneurysms in correlation with 3D-digital subtraction angiographic images: a study of 133 aneurysms. *Stroke* 33(7):1803–1808. <https://doi.org/10.1161/01.str.0000019510.32145.a9>
110. White PM, Teasdale EM, Wardlaw JM, Easton V (2001) Intracranial aneurysms: CT angiography and MR angiography for detection prospective blinded comparison in a large patient cohort. *Radiology* 219(3):739–749. <https://doi.org/10.1148/radiology.219.3.r01ma16739>
111. Yang X, Blezek DJ, Cheng LTE, Ryan WJ, Kallmes DF, Erickson BJ (2011) Computer-aided detection of intracranial aneurysms in MR angiography. *J Digit Imaging* 24(1):86–95. <https://doi.org/10.1007/s10278-009-9254-0>
112. Shi Z, Hu B, Schoepf UJ, Savage RH, Dargis DM, Pan CW et al (2020) Artificial intelligence in the management of intracranial aneurysms: current status and future perspectives. *AJNR Am J Neuroradiol* 41(3):373–379. <https://doi.org/10.3174/ajnr.A6468>
113. Nakao T, Hanaoka S, Nomura Y, Sato I, Nemoto M, Miki S et al (2018) Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography. *J Magn Reson Imaging* 47(4):948–953. <https://doi.org/10.1002/jmri.25842>
114. Ueda D, Yamamoto A, Nishimori M, Shimono T, Doishita S, Shimazaki A et al (2019) Deep learning for MR angiography: automated detection of cerebral aneurysms. *Radiology* 290(1):187–194. <https://doi.org/10.1148/radiol.2018180901>
115. Park A, Chute C, Rajpurkar P, Lou J, Ball RL, Shpanskaya K et al (2019) Deep learning-assisted diagnosis of cerebral aneurysms using the HeadXNet model. *JAMA Netw Open* 2(6):e195600. <https://doi.org/10.1001/jamanetworkopen.2019.5600>

116. Dai X, Huang L, Qian Y, Xia S, Chong W, Liu J et al (2020) Deep learning for automated cerebral aneurysm detection on computed tomography images. *Int J Comput Assist Radiol Surg* 15(4):715–723. <https://doi.org/10.1007/s11548-020-02121-2>
117. Zeng Y, Liu X, Xiao N, Li Y, Jiang Y, Feng J et al (2019) Automatic diagnosis based on spatial information fusion feature for intracranial aneurysm. *IEEE Trans Med Imaging*. <https://doi.org/10.1109/TMI.2019.2951439>
118. Duan H, Huang Y, Liu L, Dai H, Chen L, Zhou L (2019) Automatic detection on intracranial aneurysm from digital subtraction angiography with cascade convolutional neural networks. *Biomed Eng Online* 18(1):110. <https://doi.org/10.1186/s12938-019-0726-2>
119. Sichtermann T, Faron A, Sijben R, Teichert N, Freiherr J, Wiesmann M (2019) Deep learning-based detection of intracranial aneurysms in 3D TOF-MRA. *AJNR Am J Neuroradiol* 40(1):25–32. <https://doi.org/10.3174/ajnr.A5911>
120. Faron A, Sichtermann T, Teichert N, Luetkens JA, Keulers A, Nikoubashman O et al (2019) Performance of a deep-learning neural network to detect intracranial aneurysms from 3D TOF-MRA compared to human readers. *Clin Neuroradiol*. <https://doi.org/10.1007/s00062-019-00809-w>
121. Investigators UJ, Morita A, Kirino T, Hashi K, Aoki N, Fukuhara S et al (2012) The natural course of unruptured cerebral aneurysms in a Japanese cohort. *N Engl J Med* 366(26):2474–2482. <https://doi.org/10.1056/NEJMoa1113260>
122. Naggara ON, Lecler A, Oppenheim C, Meder J-F, Raymond J (2012) Endovascular treatment of intracranial unruptured aneurysms: a systematic review of the literature on safety with emphasis on subgroup analyses. *Radiology* 263(3):828–835. <https://doi.org/10.1148/radiol.12112114>
123. Thompson BG, Brown RD, Amin-Hanjani S, Broderick JP, Cockroft KM, Connolly ES et al (2015) Guidelines for the management of patients with unruptured intracranial aneurysms: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 46(8):2368–2400. <https://doi.org/10.1161/STR.0000000000000070>
124. Wiebers DO, Whisnant JP, Huston J, Meissner I, Brown RD, Piepgras DG et al (2003) Unruptured intracranial aneurysms: natural history, clinical outcome, and risks of surgical and endovascular treatment. *Lancet* (London, England) 362(9378):103–110. [https://doi.org/10.1016/s0140-6736\(03\)13860-3](https://doi.org/10.1016/s0140-6736(03)13860-3)
125. Murayama Y, Fujimura S, Suzuki T, Takao H (2019) Computational fluid dynamics as a risk assessment tool for aneurysm rupture. *Neurosurg Focus* 47(1):E12. <https://doi.org/10.3171/2019.4.FOCUS19189>
126. Chien A, Sayre J (2014) Morphologic and hemodynamic risk factors in ruptured aneurysms imaged before and after rupture. *AJNR Am J Neuroradiol* 35(11):2130–2135. <https://doi.org/10.3174/ajnr.A4016>
127. Cornelissen BMW, Schneiders JJ, Potters WV, van den Berg R, Velthuis BK, Rinkel GJE et al (2015) Hemodynamic differences in intracranial aneurysms before and after rupture. *AJNR Am J Neuroradiol* 36(10):1927–1933. <https://doi.org/10.3174/ajnr.A4385>
128. Sugiyama S-I, Endo H, Omodaka S, Endo T, Niizuma K, Rashad S et al (2016) Daughter sac formation related to blood inflow jet in an intracranial aneurysm. *World Neurosurg* 96:396–402. <https://doi.org/10.1016/j.wneu.2016.09.040>
129. Stember JN, Chang P, Stember DM, Liu M, Grinband J, Filippi CG et al (2019) Convolutional neural networks for the detection and measurement of cerebral aneurysms on magnetic resonance angiography. *J Digit Imaging* 32(5):808–815. <https://doi.org/10.1007/s10278-018-0162-z>
130. Podgorsak AR, Rava RA, Shiraz Bhurwani MM, Chandra AR, Davies JM, Siddiqui AH et al (2019) Automatic radiomic feature extraction using deep learning for angiographic parametric imaging of intracranial aneurysms. *J Neurointerv Surg*. <https://doi.org/10.1136/neurintsurg-2019-015214>
131. Zhao X, Gold N, Fang Y, Xu S, Zhang Y, Liu J et al (2018) Vertebral artery fusiform aneurysm geometry in predicting rupture risk. *R Soc Open Sci* 5(10):180780. <https://doi.org/10.1098/rsos.180780>
132. Liu Q, Jiang P, Jiang Y, Ge H, Li S, Jin H et al (2019) Prediction of aneurysm stability using a machine learning model based on PyRadiomics-derived morphological features. *Stroke* 50(9):2314–2321. <https://doi.org/10.1161/STROKEAHA.119.025777>
133. Kim HC, Rhim JK, Ahn JH, Park JJ, Moon JU, Hong EP et al (2019) Machine learning application for rupture risk assessment in small-sized intracranial aneurysm. *J Clin Med* 8(5):683. <https://doi.org/10.3390/jcm8050683>

134. Paliwal N, Jaiswal P, Tutino VM, Shallwani H, Davies JM, Siddiqui AH et al (2018) Outcome prediction of intracranial aneurysm treatment by flow diverters using machine learning. *Neurosurg Focus* 45(5):E7. <https://doi.org/10.3171/2018.8.FOCUS18332>
135. Liu J, Chen Y, Lan L, Lin B, Chen W, Wang M et al (2018) Prediction of rupture risk in anterior communicating artery aneurysms with a feed-forward artificial neural network. *Eur Radiol* 28(8):3268–3275. <https://doi.org/10.1007/s00330-017-5300-3>
136. Varble N, Tutino VM, Yu J, Sonig A, Siddiqui AH, Davies JM et al (2018) Shared and distinct rupture discriminants of small and large intracranial aneurysms. *Stroke* 49(4): 856–864. <https://doi.org/10.1161/STROKEAHA.117.019929>
137. Detmer FJ, Luckehe D, Mut F, Slawski M, Hirsch S, Bijlenga P et al (2019) Comparison of statistical learning approaches for cerebral aneurysm rupture assessment. *Int J Comput Assist Radiol Surg.* <https://doi.org/10.1007/s11548-019-02065-2>
138. Tanioka S, Ishida F, Yamamoto A, Shimizu S, Sakaida H, Toyoda M et al (2020) Machine learning classification of cerebral aneurysm rupture status with morphologic variables and hemodynamic parameters. *Radiol Artif Intell* 2(1):e190077. <https://doi.org/10.1148/ryai.2019190077>
139. Shi Z, Chen GZ, Mao L, Li XL, Zhou CS, Xia S et al (2021) Machine learning-based prediction of small intracranial aneurysm rupture status using CTA-derived hemodynamics: a multicenter study. *AJNR Am J Neuroradiol* 42(4):648–654. <https://doi.org/10.3174/ajnr.A7034>
140. Kim KH, Koo HW, Lee BJ, Sohn MJ (2021) Analysis of risk factors correlated with angiographic vasospasm in patients with aneurysmal subarachnoid hemorrhage using explainable predictive modeling. *J Clin Neurosci* 91:334–342. <https://doi.org/10.1016/j.jocn.2021.07.028>
141. Ramos LA, van der Steen WE, Sales Barros R, Majoe C, van den Berg R, Verbaan D et al (2019) Machine learning improves prediction of delayed cerebral ischemia in patients with subarachnoid hemorrhage. *J Neurointerv Surg.* 11(5):497–502. <https://doi.org/10.1136/neurintsurg-2018-014258>
142. Rubbert C, Patil KR, Beseoglu K, Mathys C, May R, Kaschner MG et al (2018) Prediction of outcome after aneurysmal subarachnoid hemorrhage using data from patient admission. *Eur Radiol* 28(12):4949–4958. <https://doi.org/10.1007/s00330-018-5505-0>
143. Wardlaw JM, Smith C, Dichgans M (2013) Mechanisms of sporadic cerebral small vessel disease: insights from neuroimaging. *Lancet Neurol* 12(5):483–497. [https://doi.org/10.1016/S1474-4422\(13\)70060-7](https://doi.org/10.1016/S1474-4422(13)70060-7)
144. Wardlaw JM, Smith EE, Biessels GJ, Cordonnier C, Fazekas F, Frayne R et al (2013) Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol* 12(8):822–838
145. Wardlaw JM, Smith C, Dichgans M (2019) Small vessel disease: mechanisms and clinical implications. *Lancet Neurol* 18(7):684–696. [https://doi.org/10.1016/S1474-4422\(19\)30079-1](https://doi.org/10.1016/S1474-4422(19)30079-1)
146. Debette S, Schilling S, Duperron M-G, Larsen SC, Markus HS (2019) Clinical significance of magnetic resonance imaging markers of vascular brain injury: a systematic review and meta-analysis. *JAMA Neurol* 76(1): 81–94. <https://doi.org/10.1001/jamaneurol.2018.3122>
147. Kwon HM, Lynn MJ, Turan TN, Derdeyn CP, Fiorella D, Lane BF et al (2016) Frequency, risk factors, and outcome of coexistent small vessel disease and intracranial arterial stenosis: results from the Stenting and Aggressive Medical Management for Preventing Recurrent Stroke in Intracranial Stenosis (SAMMPRIS) Trial. *JAMA Neurol* 73(1):36–42. <https://doi.org/10.1001/jamaneurol.2015.3145>
148. Pasi M, Cordonnier C (2020) Clinical relevance of cerebral small vessel diseases. *Stroke* 51(1):47–53. <https://doi.org/10.1161/STROKEAHA.119.024148>
149. Kapasi A, DeCarli C, Schneider JA (2017) Impact of multiple pathologies on the threshold for clinically overt dementia. *Acta Neuropathol* 134(2):171–186. <https://doi.org/10.1007/s00401-017-1717-7>
150. Bos D, Wolters FJ, Darweesh SKL, Vernooij MW, Wolf F, Ikram MA et al (2018) Cerebral small vessel disease and the risk of dementia: a systematic review and meta-analysis of population-based evidence. *Alzheimers Dement* 14(11):1482–1492. <https://doi.org/10.1016/j.jalz.2018.04.007>
151. Akoudad S, Wolters FJ, Viswanathan A, Bruijn RF, Lugt A, Hofman A et al (2016) Association of cerebral microbleeds with cognitive decline and dementia. *JAMA Neurol* 73(8):934–943. <https://doi.org/10.1001/jamaneurol.2016.1017>

152. Brown R, Benveniste H, Black SE, Charpak S, Dichgans M, Joutel A et al (2018) Understanding the role of the perivascular space in cerebral small vessel disease. *Cardiovasc Res* 114(11):1462–1473. <https://doi.org/10.1093/cvr/cvy113>
153. Georgakis MK, Duering M, Wardlaw JM, Dichgans M (2019) WMH and long-term outcomes in ischemic stroke: a systematic review and meta-analysis. *Neurology* 92(12): e1298–ee308. <https://doi.org/10.1212/WNL.00000000000007142>
154. Staals J, Makin SD, Doubal FN, Dennis MS, Wardlaw JM (2014) Stroke subtype, vascular risk factors, and total MRI brain small-vessel disease burden. *Neurology* 83(14): 1228–1234. <https://doi.org/10.1212/WNL.0000000000000837>
155. Xu X, Hilal S, Collinson SL, Chong EJY, Ikram MK, Venketasubramanian N et al (2015) Association of magnetic resonance imaging markers of cerebrovascular disease burden and cognition. *Stroke* 46(10): 2808–2814. <https://doi.org/10.1161/STROKEAHA.115.010700>
156. Prins ND, Scheltens P (2015) White matter hyperintensities, cognitive impairment and dementia: an update. *Nat Rev Neurol* 11(3): 157–165. <https://doi.org/10.1038/nrneurol.2015.10>
157. Fazekas F, Chawluk JB, Alavi A, Hurtig HI, Zimmerman RA (1987) MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *AJR Am J Roentgenol* 149(2): 351–356. <https://doi.org/10.2214/ajr.149.2.351>
158. Heuvel DMJ, Dam VH, Craen AJM, Admiraal-Behloul F, Es ACGM, Palm WM et al (2006) Measuring longitudinal white matter changes: comparison of a visual rating scale with a volumetric measurement. *AJNR Am J Neuroradiol* 27(4):875–878
159. Straaten ECW, Fazekas F, Rostrup E, Scheltens P, Schmidt R, Pantoni L et al (2006) Impact of white matter hyperintensities scoring method on correlations with clinical data: the LADIS study. *Stroke* 37(3): 836–840. <https://doi.org/10.1161/01.STR.0000202585.26325.74>
160. Mäntylä R, Erkinjuntti T, Salonen O, Aronen HJ, Peltonen T, Pohjasvaja T et al (1997) Variable agreement between visual rating scales for white matter hyperintensities on MRI. Comparison of 13 rating scales in a poststroke cohort. *Stroke* 28(8):1614–1623. <https://doi.org/10.1161/01.str.28.8.1614>
161. Anbeek P, Vincken KL, Osch MJP, Bisschops RHC, Grond J (2004) Probabilistic segmentation of white matter lesions in MR imaging. *NeuroImage* 21(3):1037–1044. <https://doi.org/10.1016/j.neuroimage.2003.10.012>
162. Lao Z, Shen D, Liu D, Jawad AF, Melhem ER, Launer LJ et al (2008) Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. *Acad Radiol* 15(3):300–313. <https://doi.org/10.1016/j.acra.2007.10.012>
163. Herskovits EH, Bryan RN, Yang F (2008) Automated Bayesian segmentation of microvascular white-matter lesions in the ACCORD-MIND study. *Adv Med Sci* 53(2):182–190. <https://doi.org/10.2478/v10039-008-0039-3>
164. Beare R, Srikanth V, Chen J, Phan TG, Stapleton J, Lipshut R et al (2009) Development and validation of morphological segmentation of age-related cerebral white matter hyperintensities. *NeuroImage* 47(1): 199–203. <https://doi.org/10.1016/j.neuroimage.2009.03.055>
165. Dyrby TB, Rostrup E, Baaré WFC, Straaten ECW, Barkhof F, Vrenken H et al (2008) Segmentation of age-related white matter changes in a clinical multi-center study. *NeuroImage* 41(2):335–345. <https://doi.org/10.1016/j.neuroimage.2008.02.024>
166. Jack CR, O'Brien PC, Rettman DW, Shiung MM, Xu Y, Muthupillai R et al (2001) FLAIR histogram segmentation for measurement of leukoaraiosis volume. *J Magn Reson Imaging* 14(6):668–676. <https://doi.org/10.1002/jmri.10011>
167. Admiraal-Behloul F, Heuvel DMJ, Olofsen H, Osch MJP, Grond J, Buchem MA et al (2005) Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *NeuroImage* 28(3): 607–617. <https://doi.org/10.1016/j.neuroimage.2005.06.061>
168. Seghier ML, Ramlackhansingh A, Crinion J, Leff AP, Price CJ (2008) Lesion identification using unified segmentation-normalisation models and fuzzy clustering. *NeuroImage* 41(4):1253–1266. <https://doi.org/10.1016/j.neuroimage.2008.03.028>
169. Wang Y, Catindig JA, Hilal S, Soon HW, Ting E, Wong TY et al (2012) Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts. *NeuroImage* 60(4):2379–2388. <https://doi.org/10.1016/j.neuroimage.2012.02.034>
170. Jeon S, Yoon U, Park J-S, Seo SW, Kim J-H, Kim ST et al (2011) Fully automated pipeline for quantification and localization of white

- matter hyperintensity in brain magnetic resonance image. *Int J Imaging Syst Technol* 21(2):193–200. <https://doi.org/10.1002/ima.20277>
171. Caligiuri ME, Perrotta P, Augimeri A, Rocca F, Quattrone A, Cherubini A (2015) Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: a review. *Neuroinformatics* 13(3):261–276. <https://doi.org/10.1007/s12021-015-9260-y>
 172. Kuijf HJ, Biesbroek JM, De Bresser J, Heinen R, Andermatt S, Bento M et al (2019) Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE Trans Med Imaging* 38(11): 2556–2568. <https://doi.org/10.1109/TMI.2019.2905770>
 173. Li H, Jiang G, Zhang J, Wang R, Wang Z, Zheng W-S et al (2018) Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *NeuroImage* 183:650–665. <https://doi.org/10.1016/j.neuroimage.2018.07.005>
 174. Andermatt S, Pezold S, Cattin P (2016) Multi-dimensional gated recurrent units for the segmentation of biomedical 3D-data. https://doi.org/10.1007/978-3-319-46976-8_15
 175. Ghafoorian M, Karssemeijer N, Heskes T, Uden IWM, Sanchez CI, Litjens G et al (2017) Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Sci Rep* 7(1):5110. <https://doi.org/10.1038/s41598-017-05300-5>
 176. Valverde S, Cabezas M, Roura E, González-Villà S, Pareto D, Vilanova JC et al (2017) Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage* 155:159–168. <https://doi.org/10.1016/j.neuroimage.2017.04.034>
 177. Zhang Z, Powell K, Yin C, Cao S, Gonzalez D, Hannawi Y et al (2021) Brain atlas guided attention U-net for white matter hyperintensity segmentation. *AMIA Jt Summits Transl Sci Proc* 2021:663–671
 178. Park G, Hong J, Duffy BA, Lee J-M, Kim H (2021) White matter hyperintensities segmentation using the ensemble U-net with multi-scale highlighting foregrounds. *NeuroImage* 237:118140. <https://doi.org/10.1016/j.neuroimage.2021.118140>
 179. Balakrishnan R, Valdés Hernández MC, Farrell AJ (2021) Automatic segmentation of white matter hyperintensities from brain magnetic resonance images in the era of deep learning and big data – a systematic review. *Comput Med Imaging Graph* 88: 101867. <https://doi.org/10.1016/j.compmedimag.2021.101867>
 180. Jiang J, Liu T, Zhu W, Koncz R, Liu H, Lee T et al (2018) UBO detector – a cluster-based, fully automated pipeline for extracting white matter hyperintensities. *NeuroImage* 174: 539–549. <https://doi.org/10.1016/j.neuroimage.2018.03.050>
 181. Sundaresan V, Zamboni G, Le Heron C, Rothwell PM, Husain M, Battaglini M et al (2019) Automated lesion segmentation with BIANCA: impact of population-level features, classification algorithm and locally adaptive thresholding. *NeuroImage* 202:116056. <https://doi.org/10.1016/j.neuroimage.2019.116056>
 182. Zhan T, Yu R, Zheng Y, Zhan Y, Xiao L, Wei Z (2017) Multimodal spatial-based segmentation framework for white matter lesions in multi-sequence magnetic resonance images. *Biomed Signal Process Control* 31:52–62. <https://doi.org/10.1016/j.bspc.2016.06.016>
 183. Ding T, Cohen AD, O'Connor EE, Karim HT, Crainiceanu A, Muschelli J et al (2020) An improved algorithm of white matter hyperintensity detection in elderly adults. *Neuroimage Clin* 25:102151. <https://doi.org/10.1016/j.nicl.2019.102151>
 184. Zhan T, Zhan Y, Liu Z, Xiao L, Wei Z (2015) Automatic method for white matter lesion segmentation based on T1-fluid-attenuated inversion recovery images. *IET Comput Vis* 9(4):447–455. <https://doi.org/10.1049/iet-cvi.2014.0121>
 185. Valverde S, Oliver A, Roura E, González-Villà S, Pareto D, Vilanova JC et al (2017) Automated tissue segmentation of MR brain images in the presence of white matter lesions. *Med Image Anal* 35:446–457. <https://doi.org/10.1016/j.media.2016.08.014>
 186. Fiford CM, Sudre CH, Pemberton H, Walsh P, Manning E, Malone IB et al (2020) Automated White matter hyperintensity segmentation using Bayesian model selection: assessment and correlations with cognitive change. *Neuroinformatics* 18(3):429–449. <https://doi.org/10.1007/s12021-019-09439-6>
 187. Damangir S, Westman E, Simmons A, Vrenken H, Wahlund L-O, Spulber G (2017) Reproducible segmentation of white matter hyperintensities using a new statistical definition. *MAGMA* 30(3):227–237.

- <https://doi.org/10.1007/s10334-016-0599-3>
188. Vanderbecq Q, Xu E, Stroer S, Couvy-Duchesne B, Diaz Melo M, Dormont D et al (2020) Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients. *Neuroimage Clin* 27:102357. <https://doi.org/10.1016/j.neuroimage.2020.102357>
189. Greenberg SM, Vernooij MW, Cordonnier C, Viswanathan A, Al-Shahi Salman R, Warach S et al (2009) Cerebral microbleeds: a guide to detection and interpretation. *Lancet Neurol* 8(2):165–174. [https://doi.org/10.1016/S1474-4422\(09\)70013-4](https://doi.org/10.1016/S1474-4422(09)70013-4)
190. Charidimou A, Shoamanesh A, Wilson D, Gang Q, Fox Z, Jager HR et al (2015) Cerebral microbleeds and postthrombolysis intracerebral hemorrhage risk updated meta-analysis. *Neurology* 85(11):927–924. <https://doi.org/10.1212/WNL.0000000000001923>
191. Wilson D, Ambler G, Shakeshaft C, Brown MM, Charidimou A, Al-Shahi Salman R et al (2018) Cerebral microbleeds and intracranial haemorrhage risk in patients anticoagulated for atrial fibrillation after acute ischaemic stroke or transient ischaemic attack (CROMIS-2): a multicentre observational cohort study. *Lancet Neurol* 17(6):539–547. [https://doi.org/10.1016/S1474-4422\(18\)30145-5](https://doi.org/10.1016/S1474-4422(18)30145-5)
192. Fazekas F, Kleinert R, Roob G, Kleinert G, Kapeller P, Schmidt R et al (1999) Histopathologic analysis of foci of signal loss on gradient-echo T2*-weighted MR images in patients with spontaneous intracerebral hemorrhage: evidence of microangiopathy-related microbleeds. *AJNR Am J Neuroradiol* 20(4):637–642
193. Knudsen KA, Rosand J, Karluk D, Greenberg SM (2001) Clinical diagnosis of cerebral amyloid angiopathy: validation of the Boston criteria. *Neurology* 56(4):537–539. <https://doi.org/10.1212/wnl.56.4.537>
194. van den Heuvel TL, van der Eerden AW, Manniesing R, Ghafoorian M, Tan T, Andriessen TM et al (2016) Automated detection of cerebral microbleeds in patients with traumatic brain injury. *Neuroimage Clin* 12:241–251. <https://doi.org/10.1016/j.neuroimage.2016.07.002>
195. Wang S, Tang C, Sun J, Zhang Y (2019) Cerebral micro-bleeding detection based on densely connected neural network. *Front Neurosci* 13:422. <https://doi.org/10.3389/fnins.2019.00422>
196. Liu S, Utriainen D, Chai C, Chen Y, Wang L, Sethi SK et al (2019) Cerebral microbleed detection using susceptibility weighted imaging and deep learning. *NeuroImage* 198:271–282. <https://doi.org/10.1016/j.neuroimage.2019.05.046>
197. Rashid T, Abdulkadir A, Nasrallah IM, Ware JB, Spincemaille P, Romero JR et al (2021) DEEPMir: a DEEP convolutional neural network for differential detection of cerebral microbleeds and iron deposits in MRI. *Sci Rep* 11(1):14124
198. Dou Q, Chen H, Yu L, Zhao L, Qin J, Wang D et al (2016) Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. *IEEE Trans Med Imaging* 35(5):1182–1195. <https://doi.org/10.1109/TMI.2016.2528129>
199. Loos CMJ, Makin SDJ, Staals J, Dennis MS, Oostenbrugge RJ, Wardlaw JM (2018) Long-term morphological changes of symptomatic lacunar infarcts and surrounding White matter on structural magnetic resonance imaging. *Stroke* 49(5):1183–1188. <https://doi.org/10.1161/STROKEAHA.117.020495>
200. Duering M, Csanadi E, Gesierich B, Jouvent E, Hervé D, Seiler S et al (2013) Incident lacunes preferentially localize to the edge of white matter hyperintensities: insights into the pathophysiology of cerebral small vessel disease. *Brain J Neurol* 136(Pt 9):2717–2726. <https://doi.org/10.1093/brain/awt184>
201. Uchiyama Y, Asano T, Kato H, Hara T, Kanematsu M, Hoshi H et al (2012) Computer-aided diagnosis for detection of lacunar infarcts on MR images: ROC analysis of radiologists' performance. *J Digit Imaging* 25(4):497–503. <https://doi.org/10.1007/s10278-011-9444-4>
202. Ghafoorian M, Karssemeijer N, Heskes T, Bergkamp M, Wissink J, Obels J et al (2017) Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin. *Neuroimage Clin* 14:391–399. <https://doi.org/10.1016/j.neuroimage.2017.01.033>
203. Jessen NA, Munk ASF, Lundgaard I, Nedergaard M (2015) The glymphatic system – a Beginner's guide. *Neurochem Res* 40(12):2583–2599. <https://doi.org/10.1007/s11064-015-1581-6>
204. Ding J, Sigurðsson S, Jónsson PV, Eiríksdóttir G, Charidimou A, Lopez OL et al (2017) Large perivascular spaces visible on magnetic resonance imaging, cerebral small vessel disease progression, and risk of

- dementia: the age, gene/environment susceptibility-Reykjavik study. *JAMA Neurol* 74(9):1105–1112. <https://doi.org/10.1001/jamaneurol.2017.1397>
205. Charidimou A, Martinez-Ramirez S, Reijmer YD, Oliveira-Filho J, Lauer A, Roongpiboonsopit D et al (2016) Total magnetic resonance imaging burden of small vessel disease in cerebral amyloid angiopathy: an imaging-pathologic study of concept validation. *JAMA Neurol* 73(8):994–1001. <https://doi.org/10.1001/jamaneurol.2016.0832>
206. Park SH, Zong X, Gao Y, Lin W, Shen D (2016) Segmentation of perivascular spaces in 7T MR image using auto-context model with orientation-normalized features. *NeuroImage* 134:223–235. <https://doi.org/10.1016/j.neuroimage.2016.03.076>
207. Ballerini L, Lovreglio R, Valdés Hernández MDC, Ramirez J, MacIntosh BJ, Black SE et al (2018) Perivascular spaces segmentation in brain MRI using optimal 3D filtering. *Sci Rep* 8(1):2132. <https://doi.org/10.1038/s41598-018-19781-5>
208. Dubost F, Yilmaz P, Adams H, Bortsova G, Ikram MA, Niessen W et al (2019) Enlarged perivascular spaces in brain MRI: automated quantification in four regions. *NeuroImage* 185:534–544. <https://doi.org/10.1016/j.neuroimage.2018.10.026>
209. Shi Y, Wardlaw JM (2016) Update on cerebral small vessel disease: a dynamic whole-brain disease. *Stroke Vasc Neurol* 1(3):83–92. <https://doi.org/10.1136/svn-2016-000035>
210. Biesbroek JM, Weaver NA, Biessels GJ (2017) Lesion location and cognitive impact of cerebral small vessel disease. *Clin Sci (Lond)* 131(8):715–728. <https://doi.org/10.1042/CS20160452>
211. Duan Y, Shan W, Liu L, Wang Q, Wu Z, Liu P et al (2020) Primary categorizing and masking cerebral small vessel disease based on “deep learning system”. *Front Neuroinform* 14:17
212. Dickie DA, Valdés Hernández MDC, Makin SD, Staals J, Wiseman SJ, Bastin ME et al (2018) The brain health index: towards a combined measure of neurovascular and neurodegenerative structural brain injury. *Int J Stroke* 13(8):849–856. <https://doi.org/10.1177/1747493018770222>
213. Jokinen H, Koikkalainen J, Laakso HM, Melkas S, Nieminen T, Brander A et al (2020) Global burden of small vessel disease-related brain changes on MRI predicts cognitive and functional decline. *Stroke* 51(1):170–178. <https://doi.org/10.1161/STROKEAHA.119.026170>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 30

The Role of Artificial Intelligence in Neuro-oncology Imaging

Jennifer Soun, Lu-Aung Yosuke Masudathaya, Arabdha Biswas, and Daniel S. Chow

Abstract

Diagnostic imaging is widely used to assess, characterize, and monitor brain tumors. However, there remain several challenges in each of these categories due to the heterogeneous nature of these tumors. This may include variations in tumor biology that relate to variable degrees of cellular proliferation, invasion, and necrosis that in turn have different imaging manifestations. These variations have created challenges for tumor assessment, including segmentation, surveillance, and molecular characterizations. Although several rule-based approaches have been implemented that relates to tumor size and appearance, these methods inherently distill the rich amount of tumor imaging data into a limited number of variables. Approaches in artificial intelligence, machine learning, and deep learning have been increasingly leveraged to computer vision tasks, including tumor imaging, given their effectiveness for solving image-based challenges. This objective of this chapter is to summarize some of these advances in the field of tumor imaging.

Key words Brain tumors, Radiogenomics, Tumor segmentation, Response Assessment in Neuro-Oncology (RANO), Response Evaluation Criteria in Solid Tumors (RECIST)

1 Introduction

With the recent emergence of artificial intelligence in neuroimaging, there is great interest in harnessing the power of new computational approaches that are inherently quantitative to non-invasively measure and classify features of brain tumors on routine and advanced magnetic resonance imaging (MRIs). Artificial intelligence (AI), including both machine learning (ML) and deep learning (DL), has the potential to automatically detect patterns in images that remain elusive to the eye of a neuroimager and to surpass human-level performance in the prediction of glioma genetics, treatment response, and long-term outcome. Theoretically, these features of AI may enable clinicians to provide greater value to the patient by allowing for expedited and more tailored treatments. This chapter will provide a brief review of primary brain

tumor epidemiology with emphasis on gliomas, evaluate present challenges in brain tumor imaging, and describe potential applications for AI.

2 Brain Tumor Epidemiology

Primary central nervous system (CNS) tumors are a rare form of cancer, with an incidence rate in adults estimated to be 23.8 per 100,000 persons [1] (*see Box 1*) [2]. However, while these tumors are rare, they constitute a significant fraction of cancer morbidity and mortality. Within the United States, approximately 10 per 100,000 are diagnosed with a primary brain tumor each year, and 6 to 7 per 100,000 are diagnosed with a primary malignant brain tumor [3]. Brain cancer incidence is the highest in Europe (age-standardized incidence rate [ASR]: 5.5 per 100,000 persons) and North America (ASR: 5.3 per 100,000 persons), along with Australia and Western Asia [3, 4]. With regard to tumor types, astrocytomas and gliomas are the second most common malignant brain tumor in adults following metastasis, and gliomas represent approximately 30% of brain tumors and 80% of all primary malignant brain tumors [4]. Gliomas vary in histology from potentially surgically curable grade 1 tumors (e.g., pilocytic astrocytoma) to aggressive grade 4 tumors (e.g., glioblastoma, GBM) with a high risk of recurrence and/or progression [5]. Accurately classifying and characterizing tumors is vital to diagnosing tumors and producing precise prognostication.

Cancer mortality is dependent on subtype and staging, and survival time after diagnosis varies greatly by grade [6, 7]. Gliomas are classified and graded based on histological and molecular markers [6, 7]. GBM is a subtype of glioma which arises from normal glial cells and consists of a group of genetically and phenotypically heterogeneous tumors [7, 8]. GBM is the most common primary CNS tumor in adults, with an incidence of 3.2 per 100,000 adults each year in Europe and America [9]. The incidence increases significantly with age, with a mean age of diagnosis at 64 for primary GBM and a peak incidence of 15.2 cases per 100,000 between the ages of 75 and 84 [9]. GBM occurrence has been associated with several genetic diseases, including tuberous sclerosis, neurofibromatosis type I, and Li-Fraumeni syndrome; however, less than 20% of patients with GBM have a strong family history of cancer, and the only well-established environmental risk factor is exposure to ionizing radiation [10]. GBM has the poorest overall survival among gliomas, with 0.05–4.7% patient survival after 5 years of diagnosis in the United States from 1995 to 2010 (95% CI 4.4–5.0) [4, 11]. Overall, mortality and prognosis vary tremendously depending on grade and subtype, and methods to more accurately predict these factors would help improve treatment and outcomes.

Box 1 Main Primary Central Nervous System Tumors

Malignant	
Astrocytomas	20–25%
Oligodendrogiomas	1–2%
Ependymal tumors	<2%
Other	8%
Non-malignant	
Meningiomas	37%
Pituitary	16%
Nerve sheath	8%
Other	7%

GBM remains one of the most lethal malignant solid tumors. The 1-year overall survival of newly diagnosed GBM is 17–30% with a 5-year survival rate of less than 5% [6]. Surgical resection followed by chemotherapy and radiotherapy remains the cornerstone treatment choice for GBM. However, the response to chemotherapy is variable, and nearly all patients suffer from recurrent disease [4]. Additionally, these tumors most frequently arise within the frontal lobe, leading to both cognitive and motor disabilities that result in loss of independence in many patients. Increasingly, molecular markers are being used for glioma classification and characterization. Mutations such as IDH1 can be a strong predictor of favorable prognosis and can assist in distinguishing among glioma subtypes [12]. Characterizing certain genetic features such as IDH1 status can aid in more accurate diagnoses and prognostication.

3 Present Challenges with Brain Tumor Imaging

3.1 Segmentation

While there have been significant advances in neuro-oncology imaging, there remain several challenges in providing accurate measurements of brain tumors. For example, a present limitation is that commonly used techniques to monitor tumor size use unidimensional and bidimensional manual measurements. While this may work for solid tumors that have a more spherical shape, the postsurgical cavity and tumors themselves of neuro-oncology patients tend to be highly irregular in shape, which increases the difficulty in obtaining accurate measurements. This stems from the fact that GBMs themselves and their recurrence commonly

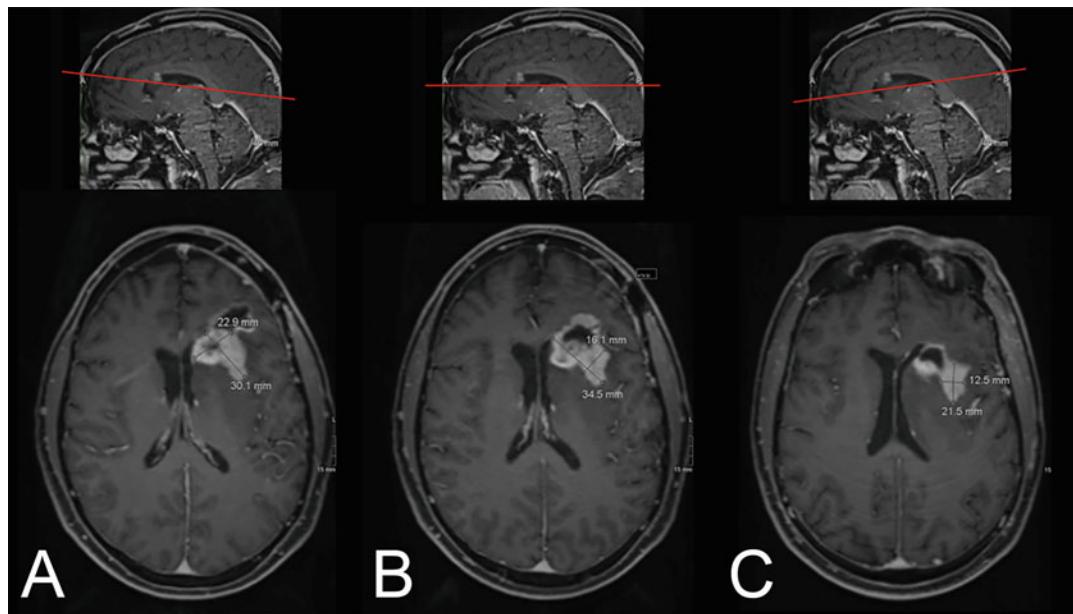


Fig. 1 Head tilt affecting designation. Patient with glioblastoma after resection. Simulation of tilting the patient's head up results in progression of disease (a) while in routine positioning demonstrates stable disease (b), and tilting downward results in partial response (c)

demonstrate eccentric and nodular growth. For patients, such inconsistencies and potential inaccuracies may result in classifying effective treatments as ineffective or vice versa (Fig. 1). Ultimately, this challenge heightens the importance for the need for reliable and reproducible techniques for tumor size measurements.

3.2 Surveillance

In addition to tumor segmentation, radiographic assessment has served as an essential tool to monitor patients with brain tumors and has played an important role in clinical trials. Historically, increases and decreases in tumor size using gadolinium contrast-enhanced sequences have served as imaging markers for progression and treatment response, respectively [13, 14]. However, there are limitations of relying solely on contrast enhancement for assessing disease status. Specifically, treatment-related increases in enhancement were observed to mimic progression with increasing frequency following the introduction of standard of care therapy of radiation and temozolomide (TMZ) [15]. This tumor pseudoprogression (psPD) is observed in 20–60% of patients who have undergone radiotherapy with TMZ and defined as increases in edema and contrast enhancement on MRI with or without clinical deterioration that subsequently stabilizes or resolves (Fig. 2) [15–17]. Additionally, the incidence has been reported to be as high as 90% in patients that have increased sensitivity to TMZ, identified with methylation status of the methyltransferase (MGMT) promoter in glioma cells [18].

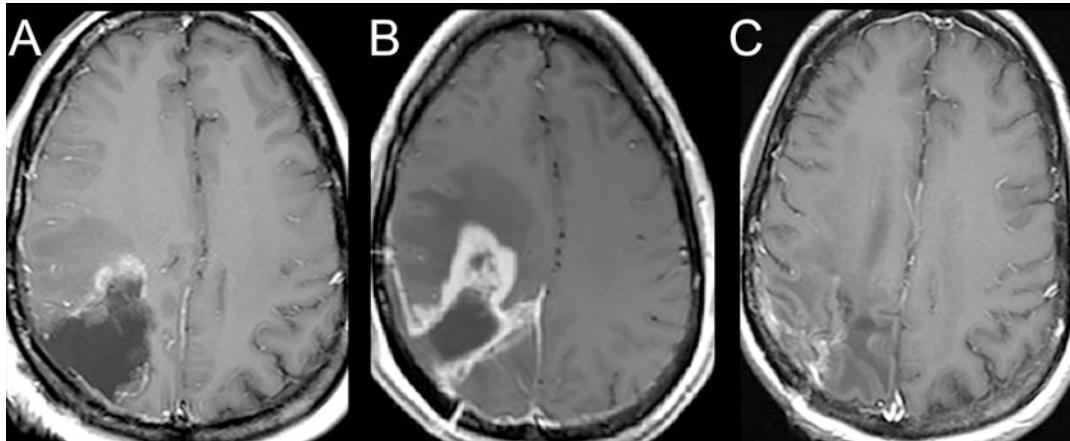


Fig. 2 Pseudoprogression. Example of a 45-year-old female with GBM. Axial post-contrast images immediately after resection show minimal enhancing disease (a). Follow-up MRI at 1 month demonstrates new thick enhancement (b) that subsequently reduced on images 12 months out (c)

Presently, the exact mechanism is still not fully understood, and the only accepted standard to distinguish true progression of disease (PD) from treatment-related psPD is invasive tissue sampling or short interval imaging or clinical follow-up, which may delay and compromise management changes in an aggressive tumor [16, 17]. In 2010, the Response Assessment in Neuro-Oncology (RANO) working group set criteria to address some of these challenges, including psPD [19]. However, evaluation of psPD remains limited with conventional imaging techniques. Challenges in monitoring GBM patients due to psPD are also observed in other newer treatments, including immunotherapies [20, 21]. The immune-related response criteria working group (iRANO) has made guidelines to address challenges of radiographic worsening in order to avoid classifying effective treatments as ineffective in instances of psPD; however, the group acknowledges that future research and solutions incorporating advanced imaging are necessary to improve assessment in these patients [21, 22].

3.3 Molecular Classification

3.3.1 Impact of Glioma Inter-tumoral Heterogeneity

Glioma inter-tumoral genetic heterogeneity has been shown to impact both prognosis and response to therapy. For example, isocitrate dehydrogenase (IDH)-mutant GBMs demonstrate significantly improved survivorship compared to IDH-wild GBMs (31 months vs. 15 months) [12, 23]. Recognition of the importance of genetic information has led the World Health Organization (WHO) to place considerable emphasis on the integration of molecular markers for its classification schemes in its 2021 update, including IDH status [24]. Regarding treatment response, it is becoming increasingly evident that GBMs' differing genetic attributes also result in mixed responses [25]. One of the early

mutations discovered was O₆-methylguanine-DNA methyltransferase (MGMT) promoter silencing, which reduces tumor cells' ability to repair DNA damage from alkylating agents such as temozolamide (TMZ). Hegi et al. [26] subsequently observed that MGMT promoter methylation silencing was observed in 45% of GBM patients, who demonstrated a survival benefit when treated with a combination of TMZ and radiotherapy versus radiotherapy alone (21.7 months versus 15.3 months). It is critical that future GBM monitoring integrates imaging and genetic data in order to provide accurate prognostic information and guide personalized therapies.

3.3.2 Challenges of Personalized Therapy

Discoveries in genetic profiling have spurred the development of new targeted therapies [27] with over 140 clinical trials presently evaluating personalized or targeted therapies for GBMs alone. These therapies are tailored to exploit genetically driven therapeutic targets. However, an apparent roadblock to these individualized approaches is the growing evidence of GBM intra-tumoral heterogeneity. Patel et al. demonstrated that GBMs consist of a mixture of cells with variable gene expression profiles using single-cell RNA sequencing [28]. Likewise, Sottoriva et al. observed genome-wide variability using surgical multisampling approach from 11 GBM patients [29]. Thus, each brain tumor may reflect multiple unique tumor habitats with corresponding differences in response and resistance to therapy, challenging the identification, development, and implementation of individualized care.

3.3.3 MRI Biomarkers of Tumor Biology and Genetic Heterogeneity

Both spatial and temporal variations in genetic expression result in alterations in tumor biology, including changes in apoptosis, cellular proliferation, cellular invasion, and angiogenesis [30]. In turn, these biologic changes manifest in the heterogeneous imaging features of brain tumors, resulting in varying degrees of enhancement and edema. For example, imaging changes on contrast-enhanced MRI result from the breakdown of the blood-brain barrier and can demonstrate areas of necrosis as a marker for apoptosis. Additionally, MRI sequences based on physiology such as apparent diffusion coefficient (ADC) and perfusion imaging have been shown to relate to tumoral cellularity and angiogenesis, respectively. Furthermore, promising efforts have shown that tumors with lower cerebral blood volume (CBV) on perfusion are more likely to be IDH mutants and have longer overall survival (OS) [31, 32]. Other reports have used enhancement patterns and ADC to predict IDH status with some success [33, 34]. Currently, efforts to provide molecular classification for brain tumors based on these MRI features have had mixed results. For example, classification of IDH and MGMT mutant status has had some success; however, methods for 1p19q and EGFR have demonstrated less reproducibility [35–37]. Different mutations may have similar MRI

features, and a “single” tumor can have multiple different mutations internally. Several approaches have emerged to provide standardized visual interpretation of gliomas for tissue classification. For example, the Visually AcceSable Rembrandt Images (VASARI) feature set is a rule-based lexicon to improve the reproducibility of interpretation [38]. However, these methods rely on human visual interpretation, which is inherently subjective and prone to inter-rater variability. Ultimately, steps are needed to provide reliable and reproducible methods to accurately classify molecular subtypes *a priori*.

4 Potential Applications for Machine Learning

4.1 Segmentation

Radiographic assessment serves an important role for clinical follow-up and research trials in oncology. Currently, the RANO criteria rely on 2D measurements of the enhancing disease as well as subjective assessment of the FLAIR non-enhancing tumor, which is then used to guide treatment strategies. However, the postsurgical cavity tends to be highly irregular in shape, which may increase the difficulty in obtaining accurate and reproducible measurements. Additionally, linear measurements obtained for cystic and necrotic tumors are often overestimated [39]. Intuitively, 3D segmentation provides a more accurate method for assessing tumor size compared to linear 2D approaches and techniques [40–42]. For example, Dempsey et al. [43] observed that 3D segmentation allows for better survival prediction compared with traditional diameter-based analysis.

Deep learning, an emerging branch of artificial intelligence, has been shown to rapidly outperform other machine learning approaches’ imaging benchmarks for various computer vision tasks [44, 45], including imaging 3D segmentation tasks. For example, Zhang et al. [46] observed that a CNN approach performed significantly better than other techniques, including random forest, support vector machine (SVM, a traditional linear machine learning technique), coupled level sets, and majority voting for brain segmentation.

Since 2012, the Multimodal *Brain Tumor Image Segmentation* (BraTS) challenge has demonstrated the efficacy of deep learning approaches for tumor segmentation [47]. This unique dataset provides developers access to GBM images, which now includes over 2000 patients from 37 institutions. As result, multiple groups have developed fully automated brain tumor segmentation tools which rely on various AI techniques to identify lesion margins and provide a more accurate estimate for disease burden (Fig. 3) [48–51]. In 2020, Isensee et al. [52] took first place with Sørensen-Dice coefficient scores of 88.95, 85.06, and 82.03 for whole tumor, tumor core, and enhancing tumor, respectively. Most recently in 2021,

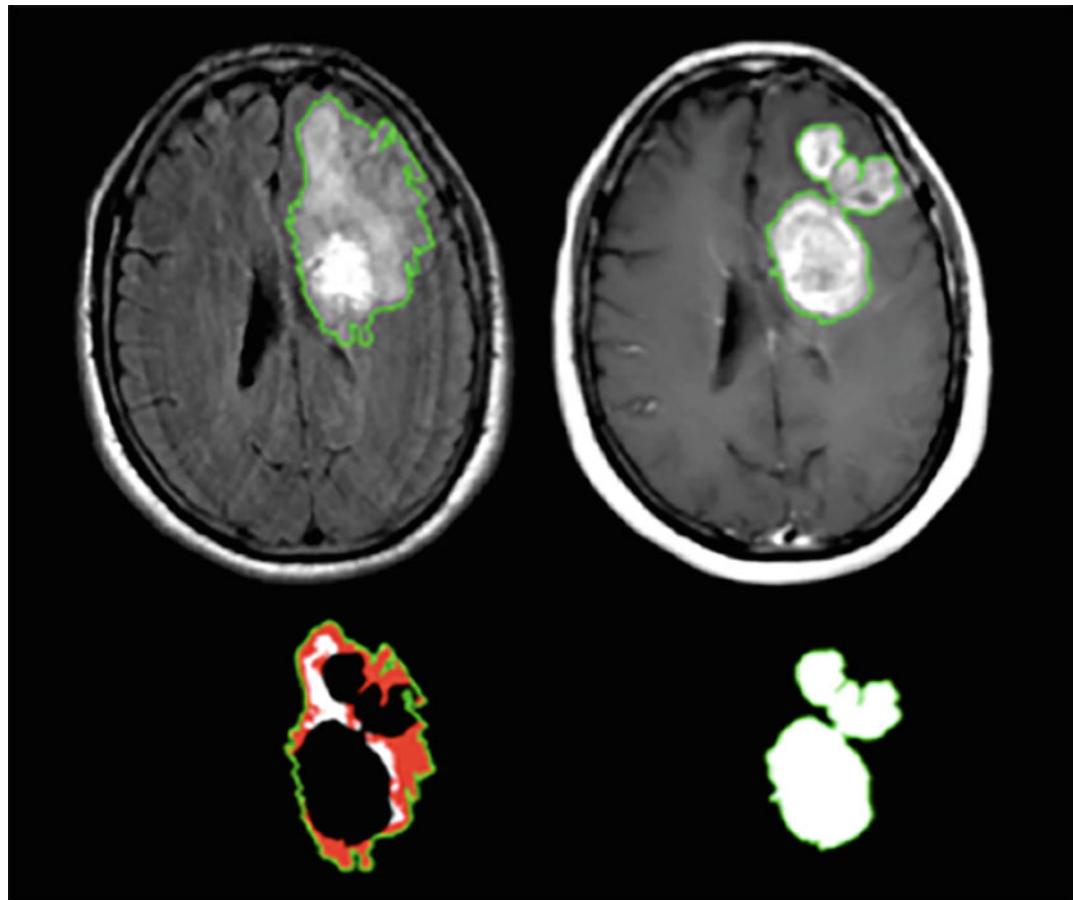


Fig. 3 Example of automated glioma segmentation using deep learning showing FLAIR edema segmentation (left) as well as segmentation of enhancing tissue (right). (Courtesy Peter Chang, MD)

BraTS has partnered with the Radiological Society of North America (RSNA) and the American Society of Neuroradiology (ASNR) [53].

4.2 Surveillance

As described previously, psPD cases are not reliably distinguished from true progression using RANO criteria with a recent meta-analysis suggesting that upward of 36% are underdiagnosed [54]. In fact, the only accepted methods to distinguish true PD from treatment-related psPD are invasive tissue sampling and short interval clinical follow-up with imaging, which may delay and compromise disease management in an aggressive tumor [16, 17].

Traditional machine learning models have been previously utilized for psPD characterization from radiologic imaging. Hu et al.'s [55] SVM approach examining multi-parametric MRI data yielded an optimized classifier for psPD with a sensitivity of 89.9% and specificity of 93.7%. Though deep learning methods have been leveraged less frequently, they are showing promise for

characterizing psPD versus true PD [56–58]. Jang et al. [56] assessed a deep learning, a long short-term memory network combined with a CNN (CNN-LSTM), to determine psPD versus tumor PD in GBM. Their dataset consisted of clinical and MRI data from 2 institutions, with 59 patients in the training cohort and 19 patients in the testing cohort. Their CNN-LSTM structure, utilizing both clinical and MRI data, outperformed the two comparison models of CNN-LSTM with MRI data alone and a random forest structure with clinical data alone, yielding an AUC (area under the curve) of 0.83, an AUPRC (area under the precision-recall curve) of 0.87, and an F-1 score of 0.74 [56]. More recently, Lee et al. [58] also utilized a CNN-STM to distinguish PD from psPD with an accuracy range of 0.62–0.75. These examples indicate that utilization of a deep learning approach can outperform a more traditional machine learning approach in analyzing images.

4.3 Molecular Classification

Radiogenomics focuses on bridging the associations between medical imaging and gene expression data in order to aid in the understanding of underlying disease mechanisms and improve diagnostics [59]. Certain molecular and genetic alterations in tissue can be observed computationally in terms of radiological appearance, including shape and texture of tissue. Radiogenomics, which leverages the interplay between radiological and genetic features in oncology, is important to improve patient treatment decisions, and artificial intelligence has become a key player that has led to significant advancements in these areas. AI-based radiogenomics has the potential to better characterize diagnosis, prognosis, and survival prediction by detecting key features in images that identify molecular characteristics of disease.

In gliomas, one of the earliest groups that used neural networks to predict tumoral genetic subtypes from imaging features was Levner et al. [60]. In this study, features were extracted from space-frequency texture analysis on the S-transform of brain MRIs to predict MGMT promoter methylation status in newly diagnosed GBM patients. Levner's group achieved an accuracy of 87.7% across 59 patients, among which 31 patients had biopsy-confirmed MGMT promoter methylated tumors. Residual CNN methods have also been used to predict MGMT promoter methylation status [61], as well as IDH mutation status. For example, Chang et al. developed a CNN to simultaneously classify IDH1, 1p19q codeletion, and MGMT promoter methylation status with high accuracy from imaging data derived from 259 patients in the Cancer Imaging Archives dataset [35]. Chang et al. also developed a principal component analysis approach to disentangle the final feature layer and determine the most influential features for each classification (Fig. 4). These features largely overlap with what has been described in the literature by subjective visual assessment. Ryu et al. [62] evaluated glioma heterogeneity via textural analysis and

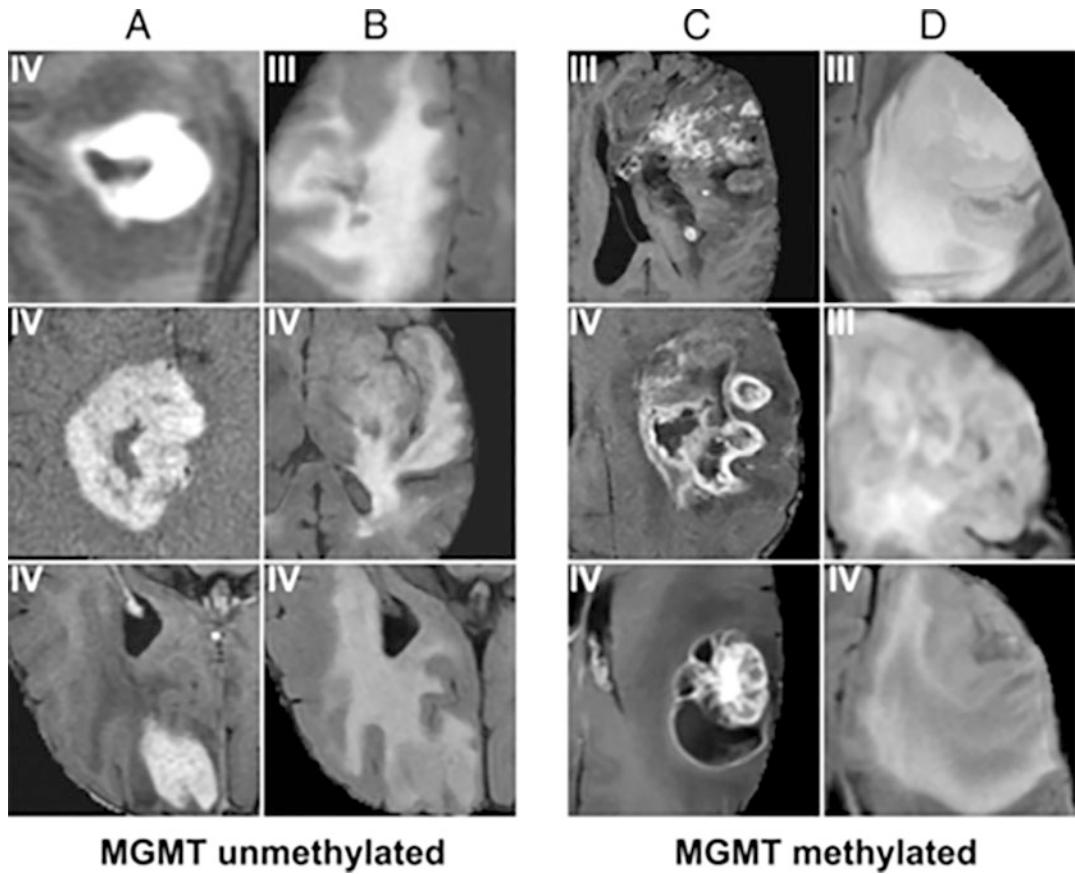


Fig. 4 MRI separating gliomas by MGMT methylation status. Features include thick enhancement with central necrosis (**a**) with infiltrative edema patterns (**b**). In contrast, features predictive of MGMT promoter methylated status include nodular and heterogeneous enhancement (**c**) with masslike FLAIR edema (**d**). (Copyright American Journal of Neuroradiology, adapted, with permission, from reference [35])

distinguished low- and high-grade gliomas with 80% accuracy. Additionally, Drabycz et al. [63] were able to classify MGMT promoter methylation status in glioblastoma patients with 71% accuracy using a textural analysis approach.

5 Summary

In summary, present challenges in brain tumor imaging in part stem from the heterogeneity of the disease, which results in challenges related to disease characterization. However, the application of novel AI, ML, and DL approaches for brain tumor imaging aims to improve many of these areas due to its ability to accurately and reliably detect imaging patterns beyond human perception. Numerous public competitions (e.g., BraTS) have also spurred the field and have recently begun collaborations with multiple

imaging societies, including the RSNA and ASNR. Ultimately, there is optimism that these tools will continue to yield new opportunities to enhance discovery and care in the future.

Acknowledgments

The authors wish to acknowledge Jack Grinband, PhD (Columbia University, New York NY); Brent Weinberg, MD PhD (Emory, Atlanta GA); and Peter Chang, MD (University of California, Irvine, Irvine CA), for their expertise and support. The authors also acknowledge the supportive administrative team from the Center for Artificial Intelligence in Diagnostic Medicine at the University of California, Irvine.

References

- Ostrom QT, Patil N, Cioffi G, Waite K, Kruchko C, Barnholtz-Sloan JS (2020) CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2013–2017. *Neuro Oncol* 22(12 Suppl 2):iv1–iv96. <https://doi.org/10.1093/neuonc/noaa200>
- Lapointe S, Perry A, Butowski NA (2018) Primary brain tumours in adults. *Lancet* 392(10145):432–446. [https://doi.org/10.1016/S0140-6736\(18\)30990-5](https://doi.org/10.1016/S0140-6736(18)30990-5)
- Wrensch M, Minn Y, Chew T, Bondy M, Berger MS (2002) Epidemiology of primary brain tumors: current concepts and review of the literature. *Neuro Oncol* 4(4):278–299. <https://doi.org/10.1093/neuonc/4.4.278>
- Ostrom QT, Gittleman H, Stetson L, Virk SM, Barnholtz-Sloan JS (2015) Epidemiology of gliomas. *Cancer Treat Res* 163:1–14. https://doi.org/10.1007/978-3-319-12048-5_1
- McNeill KA (2016) Epidemiology of brain tumors. *Neurol Clin* 34(4):981–998. <https://doi.org/10.1016/j.ncl.2016.06.014>
- Kayabolen A, Yilmaz E, Bagci-Onder T (2021) IDH mutations in Glioma: double-edged sword in clinical applications? *Biomedicines* 9(7):799. <https://doi.org/10.3390/biomedicines9070799>
- Louis DN, Perry A, Wesseling P et al (2021) The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro Oncol* 23(8):1231–1251. <https://doi.org/10.1093/neuonc/noab106>
- Urbańska K, Sokołowska J, Szmidt M, Sysa P (2014) Glioblastoma multiforme - an overview. *Contemp Oncol (Pozn)* 18(5):307–312. <https://doi.org/10.5114/wo.2014.40559>
- Ostrom QT, Gittleman H, Xu J et al (2016) CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2009–2013. *Neuro Oncol* 18(suppl_5):v1–v75. <https://doi.org/10.1093/neuonc/now207>
- Braganza MZ, Kitahara CM, Berrington de González A, Inskip PD, Johnson KJ, Rajaraman P (2012) Ionizing radiation and the risk of brain and central nervous system tumors: a systematic review. *Neuro-Oncology* 14(11):1316–1324. <https://doi.org/10.1093/neuonc/nos208>
- Ostrom QT, Bauchet L, Davis FG et al (Jul 2014) The epidemiology of glioma in adults: a “state of the science” review. *Neuro-Oncology* 16(7):896–913. <https://doi.org/10.1093/neuonc/nou087>
- Nobusawa S, Watanabe T, Kleihues P, Ohgaki H (2009) IDH1 mutations as molecular signature and predictive factor of secondary glioblastomas. *Clin Cancer Res* 15(19):6002–6007. <https://doi.org/10.1158/1078-0432.CCR-09-0715>
- Reardon DA, Galanis E, DeGroot JF et al (2011) Clinical trial end points for high-grade glioma: the evolving landscape. *Neuro-Oncology* 13(3):353–361. <https://doi.org/10.1093/neuonc/noq203>
- Macdonald DR, Cascino TL, Schold SC Jr, Cairncross JG (Jul 1990) Response criteria for phase II studies of supratentorial malignant glioma. *J Clin Oncol* 8(7):1277–1280. <https://doi.org/10.1200/jco.1990.8.7.1277>
- de Wit MC, de Bruin HG, Eijkenboom W, Sillevius Smitt PA, van den Bent MJ (2004) Immediate post-radiotherapy changes in

- malignant glioma can mimic tumor progression. *Neurology* 63(3):535–537
16. Brandsma D, Stalpers L, Taal W, Sminia P, van den Bent MJ (2008) Clinical features, mechanisms, and management of pseudoprogression in malignant gliomas. *Lancet Oncol* 9(5):453–461. [https://doi.org/10.1016/S1470-2045\(08\)70125-6](https://doi.org/10.1016/S1470-2045(08)70125-6)
 17. Hygino da Cruz LC Jr, Rodriguez I, Domingues RC, Gasparetto EL, Sorensen AG (2011) Pseudoprogression and pseudoresponse: imaging challenges in the assessment of posttreatment glioma. *AJNR Am J Neuroradiol* 32(11):1978–1985. <https://doi.org/10.3174/ajnr.A2397>
 18. Brandes AA, Franceschi E, Tosoni A et al (2008) MGMT promoter methylation status can predict the incidence and outcome of pseudoprogression after concomitant radiochemotherapy in newly diagnosed glioblastoma patients. *J Clin Oncol* 26(13):2192–2197. <https://doi.org/10.1200/JCO.2007.14.8163>
 19. Wen PY, Macdonald DR, Reardon DA et al (2010) Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J Clin Oncol* 28(11):1963–1972. <https://doi.org/10.1200/JCO.2009.26.3541>
 20. Huang RY, Wen PY (Nov 2016) Response assessment in neuro-oncology criteria and clinical endpoints. *Magn Reson Imaging Clin N Am* 24(4):705–718. <https://doi.org/10.1016/j.mric.2016.06.003>
 21. Huang RY, Neagu MR, Reardon DA, Wen PY (2015) Pitfalls in the neuroimaging of glioblastoma in the era of antiangiogenic and immuno/targeted therapy - detecting illusive disease, defining response. *Front Neurol* 6:33. <https://doi.org/10.3389/fneur.2015.00033>
 22. Okada H, Weller M, Huang R et al (Nov 2015) Immunotherapy response assessment in neuro-oncology: a report of the RANO working group. *Lancet Oncol* 16(15):e534–e542. [https://doi.org/10.1016/S1470-2045\(15\)00088-1](https://doi.org/10.1016/S1470-2045(15)00088-1)
 23. Yan H, Parsons DW, Jin G et al (2009) IDH1 and IDH2 mutations in gliomas. *N Engl J Med* 360(8):765–773. <https://doi.org/10.1056/NEJMoa0808710>
 24. Louis DN, Perry A, Reifenberger G et al (Jun 2016) The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol* 131(6):803–820. <https://doi.org/10.1007/s00401-016-1545-1>
 25. Bartek J Jr, Ng K, Bartek J, Fischer W, Carter B, Chen CC (Jul 2012) Key concepts in glioblastoma therapy. *J Neurol Neurosurg Psychiatry* 83(7):753–760. <https://doi.org/10.1136/jnnp-2011-300709>
 26. Hegi ME, Diserens AC, Gorlia T et al (2005) MGMT gene silencing and benefit from temozolamide in glioblastoma. *N Engl J Med* 352(10):997–1003. <https://doi.org/10.1056/NEJMoa043331>
 27. Auffinger B, Thaci B, Nigam P, Rincon E, Cheng Y, Lesniak MS (2012) New therapeutic approaches for malignant glioma: in search of the Rosetta stone. *F1000 Med Rep* 4:18. <https://doi.org/10.3410/M4-18>
 28. Patel AP, Tirosh I, Trombetta JJ et al (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344(6190):1396–1401. <https://doi.org/10.1126/science.1254257>
 29. Sottoriva A, Spiteri I, Piccirillo SG et al (2013) Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A* 110(10):4009–4014. <https://doi.org/10.1073/pnas.1219747110>
 30. Belden CJ, Valdes PA, Ran C et al (Oct 2011) Genetics of glioblastoma: a window into its imaging and histopathologic variability. *Radiographics* 31(6):1717–1740. <https://doi.org/10.1148/radio.316115512>
 31. Kickingereder P, Sahm F, Radbruch A et al (2015) IDH mutation status is associated with a distinct hypoxia/angiogenesis transcriptome signature which is non-invasively predictable with rCBV imaging in human glioma. *Sci Rep* 5:16238. <https://doi.org/10.1038/srep16238>
 32. Law M, Young RJ, Babb JS et al (2008) Gliomas: predicting time to progression or survival with cerebral blood volume measurements at dynamic susceptibility-weighted contrast-enhanced perfusion MR imaging. *Radiology* 247(2):490–498. <https://doi.org/10.1148/radiol.2472070898>
 33. Price SJ, Allinson K, Liu H et al (2017) Less invasive phenotype found in Isocitrate dehydrogenase-mutated glioblastomas than in Isocitrate dehydrogenase wild-type glioblastomas: a diffusion-tensor imaging study. *Radiology* 283(1):215–221. <https://doi.org/10.1148/radiol.2016152679>
 34. Xiong J, Tan W, Wen J et al (2016) Combination of diffusion tensor imaging and conventional MRI correlates with isocitrate dehydrogenase 1/2 mutations but not

- 1p/19q genotyping in oligodendroglial tumours. *Eur Radiol* 26(6):1705–1715. <https://doi.org/10.1007/s00330-015-4025-4>
35. Chang P, Grinband J, Weinberg BD et al (2018) Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *AJNR Am J Neuroradiol* 39(7):1201–1207. <https://doi.org/10.3174/ajnr.A5667>
 36. Zlochower A, Chow DS, Chang P, Khatri D, Boockvar JA, Filippi CG (2020) Deep learning AI applications in the imaging of glioma. *Top Magn Reson Imaging* 29(2):115. <https://doi.org/10.1097/RMR.0000000000000237>
 37. Shaver MM, Kohanteb PA, Chiou C et al (2019) Optimizing neuro-oncology imaging: a review of deep learning approaches for glioma imaging. *Cancers (Basel)* 11(6):829. <https://doi.org/10.3390/cancers11060829>
 38. Gutman DA, Cooper LA, Hwang SN et al (2013) MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. *Radiology* 267(2):560–569. <https://doi.org/10.1148/radiol.13120118>
 39. Chow DS, Qi J, Guo X et al (2014) Semiautomated volumetric measurement on postcontrast MR imaging for analysis of recurrent and residual disease in glioblastoma multiforme. *AJNR Am J Neuroradiol* 35(3):498–503. <https://doi.org/10.3174/ajnr.A3724>
 40. Sorensen AG, Patel S, Harmath C et al (2001) Comparison of diameter and perimeter methods for tumor volume calculation. *J Clin Oncol.* 19(2):551–557. <https://doi.org/10.1200/JCO.2001.19.2.551>
 41. Provenzale JM, Mancini MC (2012) Assessment of intra-observer variability in measurement of high-grade brain tumors. *J Neurooncol* 108(3):477–483. <https://doi.org/10.1007/s11060-012-0843-2>
 42. Provenzale JM, Ison C, Delong D (2009) Bidimensional measurements in brain tumors: assessment of interobserver variability. *AJR Am J Roentgenol* 193(6):W515–W522. <https://doi.org/10.2214/AJR.09.2615>
 43. Dempsey MF, Condon BR, Hadley DM (2005) Measurement of tumor "size" in recurrent malignant glioma: 1D, 2D, or 3D? *AJNR Am J Neuroradiol* 26(4):770–776
 44. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444. <https://doi.org/10.1038/nature14539>
 45. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. CoRR. abs/1312.6034
 46. Zhang W, Li R, Deng H et al (2015) Deep convolutional neural networks for multimodality isointense infant brain image segmentation. *NeuroImage* 108:214–224. <https://doi.org/10.1016/j.neuroimage.2014.12.061>
 47. Menze BH, Jakab A, Bauer S et al (2015) The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging* 34(10):1993–2024. <https://doi.org/10.1109/TMI.2014.2377694>
 48. Chang PD (2016) Fully convolutional deep residual neural networks for brain tumor segmentation. In: Crimi A, Menze B, Maier O, Reyes M, Winzeck S, Handels H (eds). *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries: second international workshop, BrainLes 2016, with the challenges on BRATS, ISLES and mTOP 2016*, Held in conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, revised selected papers. Springer International Publishing; pp 108–118
 49. Bangalore Yogananda CG, Shah BR, Vejdani-Jahromi M et al (2020) A fully automated deep learning network for brain tumor segmentation. *Tomography* 6(2):186–193. <https://doi.org/10.18383/j.tom.2019.00026>
 50. Ranjbarzadeh R, Bagherian Kasgari A, Jafarzadeh Ghoushchi S, Anari S, Naseri M, Bendecheche M (2021) Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. *Sci Rep* 11(1):10930. <https://doi.org/10.1038/s41598-021-90428-8>
 51. Havaei M, Davy A, Warde-Farley D et al (2017) Brain tumor segmentation with Deep Neural Networks. *Med Image Anal* 35:18–31. <https://doi.org/10.1016/j.media.2016.05.004>
 52. Isensee F, Jager PF, Full PM, Vollmuth P, Maier-Hein KH (2020) nnU-Net for brain tumor segmentation. *Int MICCAI*. arXiv preprint arXiv:2011.00848
 53. Baid U, Ghodasara S, Bilello M, et al (2021) The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314, 2021
 54. Abbasi AW, Westerlaan HE, Holtzman GA, Aden KM, van Laar PJ, van der Hoorn A (Sep 2018) Incidence of tumour progression and pseudoprogression in high-grade gliomas: a systematic review and meta-analysis. *Clin Neuroradiol* 28(3):401–411. <https://doi.org/10.1007/s00062-017-0584-x>

55. Hu X, Wong KK, Young GS, Guo L, Wong ST (2011) Support vector machine multiparametric MRI identification of pseudoprogression from tumor recurrence in patients with resected glioblastoma. *J Magn Reson Imaging* 33(2):296–305
56. Jang B-S, Jeon SH, Kim IH, Kim IA (2018) Prediction of pseudoprogression versus progression using machine learning algorithm in glioblastoma. *Sci Rep* 8(1):12516
57. Jang BS, Park AJ, Jeon SH et al (2020) Machine learning model to predict pseudoprogression versus progression in glioblastoma using MRI: a multi-institutional study (KROG 18–07). *Cancers (Basel)* 12(9):2706. <https://doi.org/10.3390/cancers12092706>
58. Lee J, Wang N, Turk S et al (2020) Discriminating pseudoprogression and true progression in diffuse infiltrating glioma using multiparametric MRI data through deep learning. *Sci Rep* 10(1):20331. <https://doi.org/10.1038/s41598-020-77389-0>
59. Trivizakis E, Papadakis GZ, Souglakos I et al (2020) Artificial intelligence radiogenomics for advancing precision and effectiveness in oncologic care (Review). *Int J Oncol* 57(1):43–53. <https://doi.org/10.3892/ijo.2020.5063>
60. Levner I, Drabycz S, Roldan G, De Robles P, Cairncross JG, Mitchell R (2009) Predicting MGMT methylation status of glioblastomas from MRI texture. *Med Image Comput Comput Assist Interv* 12(Pt 2):522–530. https://doi.org/10.1007/978-3-642-04271-3_64
61. Korfatis P, Kline TL, Lachance DH, Parney IF, Buckner JC, Erickson BJ (Oct 2017) Residual deep convolutional neural network predicts MGMT methylation status. *J Digit Imaging* 30(5):622–628. <https://doi.org/10.1007/s10278-017-0009-z>
62. Ryu YJ, Choi SH, Park SJ, Yun TJ, Kim JH, Sohn CH (2014) Glioma: application of whole-tumor texture analysis of diffusion-weighted imaging for the evaluation of tumor heterogeneity. *PLoS One* 9(9):e108335. <https://doi.org/10.1371/journal.pone.0108335>
63. Drabycz S, Roldán G, de Robles P et al (2010) An analysis of image texture, tumor location, and MGMT promoter methylation in glioblastoma using magnetic resonance imaging. *Neuroimage* 49(2):1398–1405. <https://doi.org/10.1016/j.neuroimage.2009.09.049>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 31

Machine Learning for Neurodevelopmental Disorders

Clara Moreau, Christine Deruelle, and Guillaume Auzias

Abstract

Neurodevelopmental disorders (NDDs) constitute a major health issue with >10% of the general worldwide population affected by at least one of these conditions—such as autism spectrum disorders (ASD) and attention deficit hyperactivity disorders (ADHD). Each NDD is particularly complex to dissect for several reasons, including a high prevalence of comorbidities and a substantial heterogeneity of the clinical presentation. At the genetic level, several thousands of genes have been identified (polygenicity), while a part of them was already involved in other psychiatric conditions (pleiotropy). Given these multiple sources of variance, gathering sufficient data for the proper application and evaluation of machine learning (ML) techniques is essential but challenging. In this chapter, we offer an overview of the ML methods most widely used to tackle NDDs' complexity—from stratification techniques to diagnosis prediction. We point out challenges specific to NDDs, such as early diagnosis, that can benefit from the recent advances in the ML field. These techniques also have the potential to delineate homogeneous subgroups of patients that would enable a refined understanding of underlying physiopathology. We finally survey a selection of recent papers that we consider as particularly representative of the opportunities offered by contemporary ML techniques applied to large open datasets or that illustrate the challenges faced by current approaches to be addressed in the near future.

Key words Neurodevelopmental disorders, Autism spectrum disorders, Attention deficit hyperactivity disorders, Machine learning, Pattern recognition, Classification, Clustering, Stratification

1 A Brief Introduction to Neurodevelopmental Disorders

Neurodevelopmental disorders (NDDs) cover a large range of pathologies. This term can be used to refer to known genetic syndromes such as fragile X syndrome or, in a much broader sense, include conditions with multifactorial etiology such as autism spectrum disorders (ASD), attention deficit hyperactivity disorders (ADHD), or developmental dyslexia. Even more broader are the definitions from the DSM-5 or the ICD-10 which also encompasses intellectual disabilities (ID), communication disorders, specific learning disorders, and motor disorders [1]. NDDs embrace defects that disturb the developmental function of the brain, which could lead to neuropsychiatric complications, learning

difficulties, language or non-verbal communication problems, or motor function disabilities. However, although there is a tight intrication between NDDs and psychiatric disorders—for whom manifestations come later in life—phenomenological categories used in the adult population do not apply consistently in NDDs. The latter are conditions for which the cause or the onset is located during gestation or birth and should be distinguished from late-onset disorders. We refer to [2–4] for a historical view of the standardized tools allowing for reliable and valid categorical distinctions, available to the community since the 2000s.

NDDs constitute a critical health problem in our society. More than 10% of the general worldwide population is affected by neurodevelopmental disorders [5]. The consequences of NDDs impact a person's lifetime, so patient management represents a major cost for society. Important healthcare advances have improved the life course of several NDDs (e.g., very low birth weight preterm infants, congenital hydrocephalus) and extended the expected lifespan of others (e.g., cystic fibrosis). The assessment and study of individuals with NDDs become thus an increasingly crucial issue. Researchers and clinicians have strongly emphasized the importance of early identification and intervention to improve the level of functioning. However, because of the high complexity intrinsic to these pathologies, we face a lot of misdiagnoses or even missed diagnoses which prevent early and effective therapeutic interventions. As an illustration, 1/5 of children diagnosed with ADHD or ASD in the population are currently misdiagnosed, which leads to a failure to get the adequate treatment or the administration of an unnecessary one.

NDDs are particularly complex to approach and to diagnose for several reasons. First, comorbidities are common in NDDs. Comorbid clinical features have been shown to be the rule rather than the exception in NDDs, adding to the complexity of proper diagnostic boundaries' delineation. Over a third of individuals with ASD meet criteria for ADHD, obsessive-compulsive disorder (OCD), disruptive behavior disorders, anxiety and mood disorders, intellectual disability, or epilepsy, inducing various diagnostic combinations [2, 6, 7]. This overlap across conditions probably originates from a shared neurological etiology. As a consequence, studies that exclude other psychiatric disorders have limited translational application because of the pathophysiological overlap between many comorbid disorders (*see Fig. 1* for an illustration of this issue).

In relation to this first issue, neurodevelopmental disorders overlap a lot in terms of etiology because of important epidemiological comorbidity and community of symptoms [8]. NDDs show indeed considerable overlap both neuropsychologically, physiologically, and genetically. For instance, the presence of certain behavioral characteristics, such as attention problems, does not

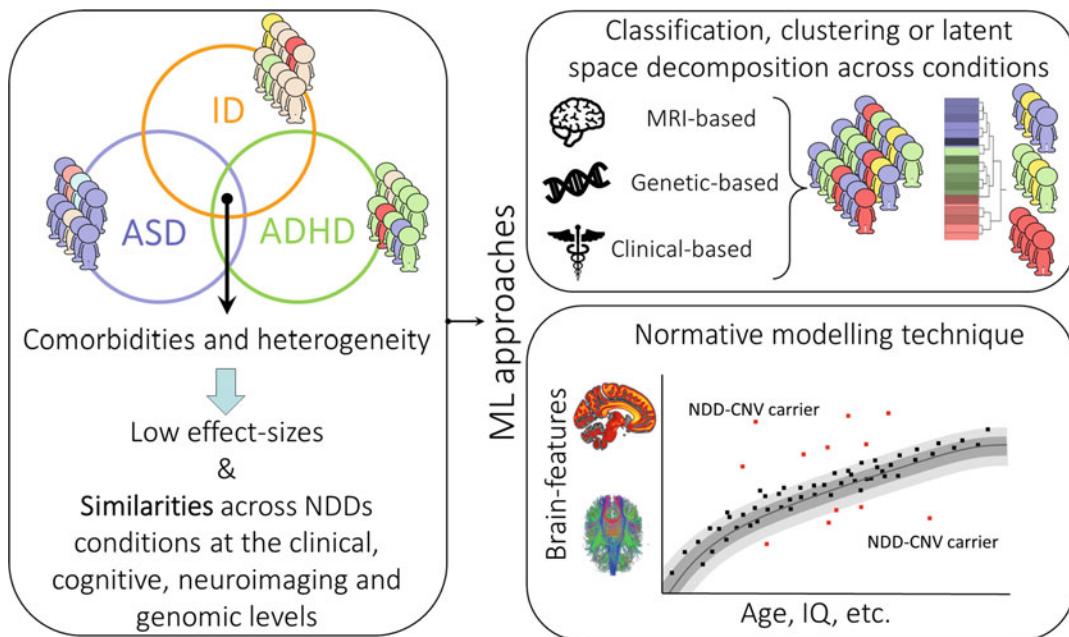


Fig. 1 *Left:* As introduced in Subheading 1, the complexity of NDDs comes from the combination of multiple sources of heterogeneity acting at different levels and that overlap across conditions as illustrated here with ASD, ADHD, and intellectual disability (ID). *Right:* As described in Subheading 2, ML approaches are instrumental to characterize and overcome the heterogeneity at each level with dedicated techniques

systematically indicate a specific diagnostic entity (e.g., ADHD), but instead, attention problems occur across a large variety of disorders (such as in ASD or in anxiety disorders). When biological bases are considered, the level of heterogeneity remains elevated. A wide range of neurological substrates have been associated with individual disorders. For example, ADHD has been associated with differences in gray matter within the anterior cingulate cortex, caudate nucleus, pallidum, striatum, cerebellum, prefrontal cortex, premotor cortex, and most parts of the parietal lobe [9].

Similarly, at the genetic level, both common and rare, and structural as well as sequence, variations have been identified as contributing to NDDs. There are multiple examples in which the identical variant has been found to contribute to a wide range of formerly distinct diagnoses, including autism, schizophrenia, epilepsy, intellectual disability, and language disorders. These include variations in chromosomal structure at 16p11.2, rare de novo point mutations at the gene *SCN2A*, and common single nucleotide polymorphism (SNP) mapping near loci encoding the genes *ITIH3*, *AS3MT*, *CACNA1C*, and *CACNB2*. In the case of autism, high genetic heritability (70–80%) with more than 1000 genes contributing to ASD has been yielded [10]. These selected examples point that heterogeneity in these pathologies is clearly multidimensional [3]. As a result, conferral of a diagnosis based on DSM-5

or ICD-10 criterion ascribes an underlying cause to the various behavioral difficulties without a method available to verify that the disorder arises from underlying biological dysfunction.

The specificity of NDDs relative to psychiatric disorders (covered in Chapter 32) is that the challenges induced by the intrication of a spectrum of conditions are potentialized by the developmental dimension. Indeed, the developmental transformation is a major contributor to the multidimensional heterogeneity across individuals affected by NDDs. Brain developmental trajectory exhibits marked variations across individuals [11, 12], but also across brain regions [13, 14]. The development course concerns cognitive, neuronal, and epigenetic maturation processes that follow distinct, yet inter-dependent, nonlinear trajectories [15, 16]. During development, reorganization and competition for function are highly active. Compensatory mechanisms can thus interfere with potential alterations of the nervous system in individuals with NDDs. The timing of these alterations is of high relevance as different neural systems are selectively vulnerable to injury at different phases of prenatal and postnatal development [17]. This plasticity partially explains the heterogeneity in behavioral and cognitive dysfunction associated with early alteration, ranging from subtle to diffuse and profound. In addition, the functional impairments can be observed immediately in some individuals, while in others, the full range of deficits may not manifest until later in life [18].

As a consequence, early diagnosis is key since early medical intervention would benefit from the remarkable plasticity of the immature brain, allowing the patient to adapt and/or develop compensatory mechanisms. On the basic research side, investigating earlier allows to reduce the influence of compensatory mechanisms and secondary perturbations. Studies focused on young children are more likely to reach the causes, whereas in adult populations, consequential or adaptation abnormalities likely contaminate the observations.

There are thus crucial needs in NDDs for a better detection of early, subtle signs of neurodevelopmental pathology and more accurate prediction of the evolution of the impairments. Gaining insight on the pathophysiological processes and the identification of more homogeneous subtypes is also required for the identification of new targets for drug development.

To address these needs, collective efforts have been made to constitute large public datasets giving access to sufficient amounts of multidimensional data covering the dimensions mentioned above (*see, e.g.,* [19]). Recently, we have witnessed the constitution of large databases trying to address these issues and which we will refer to in the following chapters. We can mention, for instance, ABCD [20], ABIDE [21], EU-AIMS [22], and ADHD200 [23] (*see Chapter 24*, for general considerations regarding the rise of openly accessible large datasets). It induced a crucial need for

statistical approaches tailored for the data-rich setting and thus called for closer collaboration with the field of machine learning.

Unsurprisingly, the NDDs having the largest prevalence, and, thus, the greater societal impact and the easier recruitment, are largely overrepresented in these databases. As a consequence, they are also overrepresented in the literature of ML techniques applied to NDDs. In the remainder of this chapter, we focus on ASD and ADHD. With regard to the characteristics mentioned above, we argue that ASD and ADHD are highly representative of the NDDs in general. As detailed in Boxes 1 and 2, they are the two most common neurodevelopmental disorders observed in childhood, and they present considerable variability, both within and across conditions. These two syndromes share most of their comorbidities, while 40–83% of children with ASD also have ADHD [24], and 28–87% of children with ASD show symptoms of ADHD [25]. See [26] for a comparison of the outcomes from recent neuroimaging studies in these two disorders. As a consequence of this heterogeneous clinical presentation, we clearly face a lack of objective criteria for diagnosis for these two disorders as well as for the other NDDs.

Box 1 Autism Spectrum Disorder (ASD)

ASD is a complex neurodevelopmental condition with life-long impacts. Current prevalence is estimated to be at least 1.5% in developed countries. The male-to-female ratio is estimated to 4:1 in this pathology. This sex ratio varies, however, according to intellectual disability (ID): reported median sex ratios of 6:1 among normal-functioning subjects and 1.7:1 among cases with moderate to severe ID [27]. Individuals with ASD suffer from a specific combination of deficits in social communication and repetitive behaviors, severely restricted interests, and sensory behaviors from early in life. Despite the vast resources devoted to the study of ASD, its pathogenesis remains largely unknown. Recent genetic studies have identified a number of rare *de novo* mutations and provided insight into polygenic risk, epigenetics, and gene-by-environment interaction related to autism or autistic traits [28]. In addition, epidemiologic investigations focusing on nongenetic factors have identified advanced parental age and preterm birth as risk factors for ASD and have suggested that prenatal exposure to air pollution and short inter-pregnancy interval are also potential risk factors. See, e.g., [29] for more detailed information.

Box 2 Attention Deficit Hyperactivity Disorder (ADHD)

ADHD is one of the most common neurodevelopmental disorders, characterized by inappropriate and developmentally harmful levels of inattention, hyperactivity, and impulsivity. It affects boys more often than girls. Its prevalence in the general population is between 3% and 4%. ADHD is diagnosed according to strictly defined criteria, but there is still no reliable biomarker of the pathology. The causes of ADHD are complex and multifactorial, with genetics, early environment, and gene-environment interplay being involved. Although ADHD is highly heritable, and multiple types of genetic variants are associated with the disease, none of them can be used as diagnostic. Diagnostic thresholds are given by both the ICD-10 and the DSM-5, but the clinical features of ADHD behave as continuously distributed dimensions and vary considerably between individuals. Clinical features are heterogeneous. ADHD profiles include not only its definite symptoms (hyperactivity-impulsiveness, inattention) and features of other neurodevelopmental disorders but also additional cognitive deficits such as impaired working memory and planning. Early comorbidity with developmental, learning, and psychiatric problems, such as ASD, is very frequent. ADHD is lifelong, but its course and outcome are highly variable. Core symptoms such as the hyperactivity observed at preschool age may turn into inattention and executive dysfunction in older children, for instance. See, e.g., [30] for further information.

2 What Are the Main Challenges in These Conditions That Can Be Addressed Using Machine Learning?

Given these multiple sources of variance, gathering sufficient amounts of data for the proper application and evaluation of machine learning (ML) techniques is essential, but also very challenging. As underlined earlier and illustrated on Fig. 1, NDDs, and more specifically the two we focus on, present a number of specific challenges that can be formulated in terms of heterogeneity, trajectory of development, and comorbidities.

In this section, we give an overview of the methods most widely used in the NDDs' literature and point to specific challenges that can benefit from the recent advances from the ML field. We refer readers interested in an exhaustive view of the available approaches and their performances in the context of NDDs to the following

recent review papers [31–36]. We organize this overview by following the historical evolution of the methods used in the field. The first applications of ML techniques were focused on classification tasks. Indeed, classification techniques can be designed for the prediction of later evolution and are thus in principle well suited to address the challenge of early diagnosis. We then observe a progressive shift toward regression, latent space decomposition, and stratification purposes. These approaches have the potential to uncover more homogeneous subpopulations of patients that would enable the refined understanding of underlying physiopathology. More recently, specific approaches have been proposed for characterizing the atypical brain maturation trajectory in NDDs. Finally, we discuss the potential of deep learning techniques for learning representations that might represent a major step toward prediction at the individual level, which is crucial for translation into clinical applications.

2.1 The Classical Analysis Approach Failed to Reach Consensus

Historically, the classical analysis approach consisted in designing a study starting from the definition of an “atypical” population of interest, based on particular clinical scores selected among the behavioral assessments used for diagnosis. This population of interest is compared to a group of control subjects, following a feature defined a priori such as “the volume of a specific cortical region estimated from anatomical MRI.” As extensively described in, e.g., [37–39], this corresponds to statistically testing the hypothesis: Does the atypical population differ, on average, from controls in the selected feature? Statistically speaking, this amounts to a case-control study using univariate hypothesis testing for one or a few features. The large literature of early studies following this approach allowed to refine the characterization of the different sources of heterogeneity presented above and shed light on the lack of biological validity of categorical representations of NDDs that manifest in the evolution of the nosology, for instance, moving from “autism” to “autism spectrum disorders” [3]. However, as we progressed in our understanding of the interactions between genetics, biological brain, and behavior, the limits of the group statistics and univariate approaches became obvious.

2.1.1 Limitations of Classical Univariate Analysis Techniques

The univariate approach is prevalent in the literature for historical reasons. It relies on the implicit assumption that different brain regions and/or different features are independent, while more and more evidence supports the opposite view: effects are spread across several brain regions, possibly located far from each other. Knowing the various sources of variance in NDDs’ data described earlier, it is unlikely that a single feature may capture a large portion of that variation and thus be interpreted in terms of underlying biological processes. It is thus not surprising that the effect sizes reported in meta-analyses remain small. In addition to potentially reduced

statistical power, the problem of inflated false discovery rate in univariate analysis framework has been raised and extensively discussed [40]. Multivariate approaches are much more relevant in this context. Indeed, combining in a multivariate approach a group of features having small effect size when considered independently might lead to a large effect [38].

2.1.2 Limitations of Group Statistics

As extensively discussed in [41], group statistics all focus on first-order statistics (group means), thereby seeking a pattern of atypicality that is consistent across the population (i.e., the “average patient”). Indeed, mean group differences may reflect a systematic shift in the distribution of the clinical group and thus provide useful information on altered processes in that population. However, those differences do not delineate variability within groups [38]. In addition, the evolution of the DSM by regrouping conditions that were considered in previous versions as distinct (e.g., Asperger and pervasive developmental disorders not otherwise specified) induced an increase in the heterogeneity of the populations included in studies on ASD [37]. Group comparisons based on diagnosis thus present the major caveat of ignoring psychiatric comorbidities, which are common in NDDs. It thus becomes obvious that group statistics applied to populations defined based on diagnostic categories are inadequate. Indeed, categorical diagnoses from the DSM are increasingly found to be incongruent with emerging neuroscientific evidence that points toward shared neurobiological dysfunction underlying NDDs [42]. See, e.g., [39] for extensive discussions on the limitations of the diagnostic-first approach in comparison to the alternative strategy that begins at the level of molecular factors enabling the study of mechanisms related to biological risk, irrespective of diagnoses or clinical manifestations.

The combination of univariate statistics and mean group difference analysis applied to heterogeneous populations with small sample sizes resulted in highly inconsistent findings. Indeed, most of the published findings are not consistent and were not replicated. The recent challenge [43] further illustrates the intrinsic limitation of the group statistics framework, but also that state-of-the-art ML techniques do not systematically outperform classical approaches in such a binary classification task. In this context, deep learning techniques were prone to overfitting with poor generalization to unseen dataset, while simpler approaches had a stable prediction performance when applied to new data. It is important to stress that several limitations from this early literature do fully apply to more advanced ML techniques and/or multivariate data analysis strategies. While the problem of inflated false discovery rate in univariate analysis framework has been extensively discussed [40], the problems related to the improper evaluation and validation of ML

techniques (e.g., overfitting and biases induced by inadapted cross-validation strategy or absence of a truly independent test set) emerge in the recent literature [31, 44–46]. While discussing the limitations of cross-validation for estimating the potential overfitting of statistical models is beyond the topic of this chapter, we stress the crucial importance of raising awareness of these aspects. We refer interested readers to essential guidelines and recommendations that have been provided in [43, 47–51]. Indeed, uncovering potential biases in the models' validation strategy is a tedious but essential step. Abraham et al. [52] is a nice illustration of the major gains in interpretation resulting from an extensive analysis of the most influential factors.

2.2 Promises of ML in NDDs

The rise of big data and the sustained advances in ML enable in principle the integration of various and heterogeneous characteristics such as behavioral profiles, imaging phenotypes, and genomics. The extraction and manual construction of features from each data type, also termed as *feature engineering*, did undergo continuous progress in tight relation with innovations in the acquisition processes. As an illustration, the imaging phenotype today covers a wide range of features extracted mainly from MRI data. For instance, a variety of measures can be extracted from diffusion-weighted imaging [53], from basic estimation in each voxel such as the *fractional anisotropy* to higher-level connectivity measures in each anatomically defined *fiber tract*, or even connections between distant anatomical regions (structural connectivity). On the genetics side, polygenic risk scores (PRS) are additive models developed to estimate the aggregate effects of thousands of common variants with very small individual effects. They can be computed for any individual to estimate the risk/probability for a particular trait conferred by common variants [54]. Feature engineering is a crucial step in the analysis since the biological relevance of the features directly impacts the interpretation, and the strategy used to manage potential interaction across different features might determine the performance of the analysis procedure more than the ML algorithm itself. In parallel, the increase in the size of the available data enables the training of more complex algorithms, making it possible to investigate central questions related to the dynamics of normal and abnormal development by means of advanced ML techniques.

2.3 Classification and Prediction: Supervised Learning for NDDs

Classification techniques consist in learning a model allowing to separate different groups of subjects based on a set of training data that have been labeled and are thus subtypes of *supervised* machine learning techniques. In this context, classification techniques integrate biological and/or behavioral measures in order to extract a *predictive pattern* corresponding to the diagnosis. Classification techniques used in the literature of NDDs are the same as those used in the field of psychiatry and span the whole range of methods

detailed in Chapters 1, 2, 3, 4, 5, and 6, from simple linear models to most recent deep networks. References [31, 32, 34, 51, 55] provide a detailed overview of the recent applications of classification techniques in the context of ASD and ADHD. The general trends indicate that linear discriminant and logistic regression classifiers were prominent until around 2014, most studies focusing on a single modality (usually structural or functional MRI). Support vector machines (SVM) then became the most commonly used approach due to their performance in the small-sample high-dimension regime but also their ability to perform nonlinear classification. Approaches based on ensembles of classifiers were more recently developed to combine data from several modalities or acquired in different settings (e.g., different scanners). Even more recently, deep learning techniques' neural networks were applied to populations of a few hundred subjects. We will discuss the potential of these advanced approaches later, in a dedicated section. In terms of input data types, structural and functional MRI modalities are overrepresented in comparison to diffusion MRI, EEG, and behavioral data. Classification techniques based on genetics are getting more and more attention (e.g., using polygenic risk scores). Due to the complex and specific data preprocessing required for each modality (*see*, e.g., [43]), combining features extracted from several modalities into a multimodal classification technique represents important additional challenges. Only a few studies did explore the potential of combining several modalities so far (e.g., 4 studies among 57 reviewed in [31]), but the initiatives for sharing preprocessed data such as those in [23, 56] will facilitate this type of analyses in the future. Multimodal classification techniques did not demonstrate major performance gain so far, but further improvements can be expected by better exploiting the complementarity of the information across different modalities [32]. In terms of classification performances, the high accuracy (>80%) reported in early studies tended to decrease, while sample size increased [31, 32], suggesting that the impressive results obtained on small cohorts were affected by overfitting, sampling biases, and artificially reduced heterogeneity within and across the populations involved. Note that the decreasing effect sizes of group comparison studies might also be related to the evolution in the definition of autism toward a more inclusive and heterogeneous population [57].

In parallel with this decrease with time in the performance, the research field on psychopathology did initiate a shift, moving away from diagnostic categories based on symptoms to the concept of dimensions related to more objective measures and having better cognitive and biological validity. In particular, the US National Institute of Mental Health initiated in 2009 the Research Domain Criteria (RDoC) project to develop a classification system for mental disorders based upon fundamental dimensions of neurobiology

and observable behavior that cut across current heterogeneous disorder categories [58, 59]. Of note, this research classification system diverges from one intended for routine clinical use in multiple respects [60]. Following this progressive conceptual shift, the major methodological challenge to be addressed moved away from classification and diagnostic prediction to latent space decomposition and stratification.

2.4 Latent Space Decomposition and Clustering: Unsupervised Learning for NDDs

Following the progressive confirmation of the inadequacy of mutually exclusive diagnostic categories, behavioral assessments for quantifying ASD traits in any given individual were introduced, such as the autism spectrum quotient questionnaire [61] and the Social Responsiveness Scale (SRS) [62]. A number of studies used these scores to demonstrate that ASD traits are also present in the typically developing population as well as in other NDDs such as ADHD [63]. These studies supported the view of a continuum across NDDs and emphasized the need for novel approaches to identify general psychopathology dimensions that cut through diagnostic boundaries. Such data-driven dimensions would ultimately enable the identification of new targets for treatment development and to stratify the NDDs in subgroups more appropriate for treatment selection [58, 59, 64]. Uncovering the hidden intrinsic structure in the data is a well-known ML problem that has been formulated as *unsupervised* learning in opposition to *supervised* learning tasks such as classification where the algorithm learns to predict a label based on a training set for which the true label is known (see Chapters 1 and 2 or, e.g., [65]). Unsupervised ML techniques consist in fitting a statistical model to the data by implementing specific assumptions regarding the relationships between the input features and on the supposed hidden structure. A general assumption to all unsupervised techniques is that there exists a non-negligible degree of correlation across some of the features in the actual data, which justifies the search for a more compact optimal representation. Depending on the assumptions regarding the hidden structure to discover, unsupervised techniques can be divided into two classes: latent space decomposition and clustering. Latent space decomposition techniques aim at projecting the data onto a new feature space of lower dimension in which a large portion of the variance can be explained by a few factors. The underlying assumption is that the projected features vary continuously along the axes of this compact subspace. In contrast, clustering techniques seek to partition the data into distinct groups (often termed as population stratification) so that the observations within each group are similar to each other, while observations in different groups differ from each other. The underlying assumption is thus that a categorical representation is more appropriate than in the case of the latent space decomposition approach. In contrast with the classification task, the algorithm is

designed in this case to identify homogeneous subpopulations within and across diagnostic categories. Several recent approaches propose a unified framework combining the advantages of both the dimensional and categorical models [3, 66, 67].

All unsupervised approaches face two main challenges in the context of NDDs. First, since we are dealing with a limited amount of data, the number of dimensions or clusters that can be identified needs to remain limited in order to avoid the curse of dimensionality, i.e., when an infinite number of solutions can fit data equally well [64, 65, 68]. As a consequence, in the majority of studies, the set of input features (and thus the dimension of the input space) is selected based on data availability or prior knowledge, which raises the problem of establishing an optimal set of variables of particular relevance for NDDs [69]. Automated feature selection procedures can be used to reduce the dimensions to be explored (*see* [69] for a recap of the approaches explored so far in ASD), but the fundamental problem of limited amount of data relative to the very large dimension to explore remains [64]. The second major challenge is the validation, since with unsupervised approaches no ground truth data is available by definition, unlike in the case of supervised ML. The relevance of the resulting dimensions or clusters should be assessed in terms of interpretability relative to external measures that would ideally have some clinical relevance. Replication on a fully independent dataset allows to assess the generalizability and reduces the risk of overfitting. This is however very hard to achieve since the number of datasets available with identical measures is limited. As a consequence, it is crucial to keep in mind that unsupervised learning is only meaningful in relation to some context [70]. As extensively discussed in [64], “due to the vast dimensionality of the human population (based on environment, behavior, biology/physiology, etc.) there are multiple ways that the population might be subcategorized that are valid and ‘real’; however, any given subgrouping might not be important for the question we care about.”

Contrary to the classification task where the literature is very rich, latent space decomposition and stratification studies in NDDs are emerging approaches, and only few findings have been published so far. Two recent publications review unsupervised approaches applied to neuroimaging in the context of ASD: [31] covered 19 studies published since 2018, and [69] identified 12 studies among which 2 were already included in [31]. For an extensive review covering the literature back to 2001, *see* [71]. The methods used range from the most common such as principal component analysis for latent space decomposition and K-means for clustering to more advanced techniques such as nonnegative matrix factorization, spectral clustering, Gaussian mixture models, and Bayesian latent factor analysis such as Indian buffet processes. Most advanced approaches such as Bayesian latent factor analysis

techniques enable to infer the number of latent factors and number of putative subpopulations from the data and can be interpreted in terms of both categorical and dimensional aspects of the heterogeneity in NDDs [69, 72]. On the genomics side, multivariate approaches such as canonical correlation analysis and partial least square regression are the tools of choice for investigating the relationship between genomic variants, neuroimaging features, psychiatric conditions, and behavioral traits [39]. The development of specific methods allowing to better model the multivariate genetic covariance structure in genome-wide association studies is a very active field. For instance, [73] introduced a new approach called genomic structural equation modeling, which allows to investigate shared genetic effects across phenotypes, while concurrently testing for causes of divergence. Importantly, this evolution in the methods reflects the progressive integration of latent space decomposition and clustering techniques into unified approaches. A promising avenue of research that benefited from access to larger datasets in the past years consists in combining neuroimaging and genomics. Indeed, the effects of latent factors derived from genomics on neuroimaging endophenotypes demonstrate higher reproducibility and larger effect size than in the previous literature [39, 74].

In terms of evaluation and performances, the studies are highly dependent on the data and the assumptions that are made, either implicitly or explicitly. An illustration of this dependency on the application is the variation in the number of subtypes reported, ranging from two to six across the neuroimaging studies on ASD included in the two reviews [31, 69]. In [71], the authors cover a much broader literature (159 articles) by relaxing inclusion criteria compared to the two others. This exhaustive review identifies seven validation strategies, defined as follows: “cross-method replication,” “subtype separation,” “independent replication,” “temporal stability,” “external validation,” “parallel validation,” and “predictive validation.” They provide the distribution of the number of identified subtypes across the reviewed studies, with a range of values varying between 1 and 16, but 82% of all studies report between two and four subtypes. Of note, this chapter underlies as major challenges the access to large and multidimensional datasets and the design of an unbiased validation framework. We refer interested readers to [71], in particular for the didactic description of the various validation strategies that apply to the literature of ASD and more generally to psychiatry or other clinical groups.

2.5 Normative Modeling for NDDs

Normative modeling gained great interest in the context of psychiatry recently, and the first applications to NDDs confirm the particular relevance of this approach in this context. Marquand et al. [75] introduced normative modeling as an alternative to clustering for parsing heterogeneity across the full range of population variation, i.e., spanning both clinical and healthy cohorts. In the approach

proposed by [75], the normative models were estimated using Gaussian process regression [76]. The flexibility of this Bayesian method enables to define a mapping between any quantitative biological measures and clinically relevant variables and offers desirable properties such as robustness to overfitting and principled ways for tuning hyper-parameters. Gaussian process regression is flexible but does not scale with an increase in sample size. More importantly, this technique can lead to inaccurate uncertainty estimates when the data are non-Gaussian [77]. Less demanding alternative approaches have been proposed. In [78], the authors used a non-parametric local weighted regression to fit a smooth curve through data points. Based on the assumption that the estimated regression is likely to be smooth, [79] proposed to estimate nonlinear effects using a smoothing spline model. This approach is a special case of Gaussian process regression. It is thus less adaptive, but presents a lower computational cost than Gaussian process regression. Fraza et al. [80] presented a novel framework based on spline interpolation combined with likelihood warping and Bayesian estimation that allows to scale normative modeling to big data cohorts. Another approach based on generalized additive models was proposed in [81, 82]. The very last version of normative models was presented recently by [83] with the generalized additive models for location, scale, and shape (GAMLSS), a flexible modeling framework that can model heteroskedasticity, nonlinear effects of variables, and hierarchical structure of the data. As demonstrated in [84] with features extracted from more than 120,000 MRI, these models can be estimated on very large datasets. They are however not suitable for small datasets since the higher flexibility of such a model would be detrimental and might lead to overfitting.

Normative models are highly relevant for analyzing neuroimaging data since they can be fit at each brain location to estimate regional specificity. In the context of NDDs, two advantages are particularly critical. First, normative modeling is efficient to disentangle the effects related to brain maturation dynamics and neurodevelopmental diseases in a data-driven way. Indeed, the Bayesian framework enables estimating distinct variance components. The effect of age within the reference cohort is estimated by nonlinear interpolation, which is appropriate in this period of highly active neurodevelopment [14, 85].

Second, normative modeling provides uncertainty measures to quantify the variation across the estimated mean within the reference cohort and the deviation of each patient from the group mean. This enables the detection and mapping of subject-specific patterns of abnormality in each individual. The statistical inference at the level of the individual participant is the key to explicitly characterize the heterogeneity underlying clinical conditions. It represents a concrete alternative to the limitations of the case-control analysis

seeking a pattern of atypicality that is consistent across the population as discussed in Subheading 2.1. In the normative modeling framework, a deviation map is computed for each individual based on extreme values statistics, which does not require that atypicalities overlap across participants. These individual deviation maps can then be analyzed (e.g., using unsupervised ML approaches described in Subheading 2.4) to identify distinct patterns of abnormality, i.e., to characterize putative subpopulations.

See [41, 83, 86, 87] for further description of the normative modeling framework and recommendations to guide future applications. The release of two python packages contributed to the widespread use of this approach: <https://github.com/ppsp-team/PyNM> and <https://github.com/amarquand/PCNtoolkit>. A didactic tutorial with a step-by-step comparison of the different normative modeling approaches on synthetic data illustrating their advantages and limitations is available online here: <https://github.com/ppsp-team/PyNM/tree/master/tutorials>.

2.6 Potential and Challenges of Deep Learning

Deep learning is a class of ML algorithms characterized by their specific internal architecture as multi-layered neural networks. These multiple layers enable the striking capacity to progressively extract higher-level features without extensive prior injection. Their advantages compared to previous approaches are of crucial importance in a large range of applications and explain the considerable attention gained by DL in the wider scientific community. See, e.g., [88] for a detailed description of the DL methods used in the literature to investigate the neuroimaging correlates of psychiatric and neurological disorders. Conceptually, DL techniques are particularly relevant for the investigation of NDDs for the following reasons:

- *Integrated learning of hierarchy of features.* As mentioned in Subheading 2.2, classical ML algorithms leverage sets of structured features extracted from the input data. This feature engineering step relies on a priori regarding the data and has a strong influence on the performances. DL algorithms process directly the raw data without requiring prior feature extraction. During the learning, the algorithm can determine the optimal hierarchy of most relevant features for representing the data, resulting in a more objective process.
- *Learning relevant spatial relationships from neuroimaging data.* In the context of neuroimaging, a striking advantage of DL is its capacity to learn relevant spatial relationships among the image domain, such as an atrophy distributed across a network of several brain regions supporting a specific function [89]. In classical ML techniques, the feature engineering step and the learning phase are dissociated, such that relevant spatial

relationships may be lost. On the contrary, this spatial relationship might be preserved by DL techniques and integrated into the optimal hierarchy of features.

- *Learning nonlinear relationships and biologically relevant compact representations.* As already discussed in Subheading 2.5, nonlinear relationships across data or dimensions relevant to NDDs are expected. Conceptually, the combination of the multiple layers available in DL architectures enables to encode this nonlinearity into a cascade of nonlinear transformations while reducing the input space into a lower-dimensional “latent space,” providing a compact representation of the data. The recent works from [89–91] demonstrated that DL can exploit the presence of nonlinearity in neuroimaging data to learn generalizable representations highly relevant for characterizing the human brain. They combined supervised and unsupervised tasks in a DL framework which consisted in learning the representation from classification tasks (predicting age and sex) and then applying decomposition and clustering techniques to the latent space. These studies strongly support that DL approaches can provide more accurate mappings of the effects of age and sex on brain MRI than simpler models. The resulting representations obtained in these works are instrumental for refining the link between cognition and underlying brain systems. Another promising avenue of research denoted as scientific machine learning (<https://sciml.ai>) consists in injecting traditional scientific mechanistic models into modern deep learning architectures in order to combine the benefits of efficient data-driven automatic learning with better interpretability and integration of biophysical constraints. See [92] for a review discussing the potential of these approaches in computational neuroscience and [93] for an example application to neuroimaging data. DL techniques can thus learn representations of data that have the potential to help explain the biological underpinnings of mental disorders, providing that enough data is available.

3 A Non-exhaustive Survey of Existing Papers on Machine Learning for NDDs and Their Limitations

We refer to the recent reviews [31, 32, 34, 51, 55, 69, 71], for a complete overview of the literature of the field. Here, we survey a selection of very recent works that we consider particularly relevant with respect to the opportunities offered by recent ML techniques applied to large open datasets, or that illustrate the challenges faced by current approaches, to be addressed in the near future.

3.1 Using ML Techniques on Neuroimaging Data to Predict the Diagnosis

An international challenge (146 challengers) has been organized to predict ASD diagnosis based on several neuroimaging modalities [43]. This challenge was conducted on the largest sample available to date (>2000 individuals from the ABIDE dataset and a second, private dataset not open to challengers). An additional dataset from the EU-AIMS project [22] was used to evaluate the reproducibility of the prediction on an independent dataset (out-of-sample prediction). The ten best submissions used either logistic regression as a first-layer predictor, linear vector classification, or a combination of different methods. Best algorithms managed to predict ASD diagnosis with an in-sample AUC of 0.80. Resting-state fMRI data was a better diagnostic predictor than anatomical MRI, and simple logistic regression performed better than complex graph convolutional deep learning models (likely due to overfitting). Finally, the performances of the best algorithms decreased to an out-of-sample AUC of 0.72 (on the external sample). Authors projected that 10,000 individuals might be necessary to reach the optimal prediction.

Another study of interest was led by the consortium “Infant Brain Imaging Study” (IBIS) [94]. The authors investigated whether infants at high familial risk for autism present early postnatal atypical brain volume. A deep learning algorithm used surface area at 6 and 12 months to successfully predict an early diagnosis of autism in infants at high risk of autism at 24 months (in-sample predictive value of 81%, no out-of-sample prediction accuracy provided). These results should be tempered by several major pitfalls. First, the diagnosis of ASD is very challenging at that early age. Second, the sample size was very small (15 high-risk infants diagnosed with autism at 24 months) and thus does not comply with the recommended practices for predictive modeling [46]. Third, the specificity of the results with respect to other NDDs was not assessed. A confirmation of the reproducibility of these results in a larger, external cohort would thus be much welcome.

Overall, these results showed that applying prediction algorithms on large enough imaging data could be instrumental for the early detection of ASD and therefore early intervention. In line with the conclusions of previous reviews [31, 69], these studies also demonstrated the relevance of using imaging data as an intermediate phenotype between the biological cause (e.g., deletion of the gene content at the 16p11.2 chromosomal segment) and the associated phenotype (e.g., ASD, ADHD, intellectual disability).

3.2 Latent Space Decomposition and Subtyping Approaches Applied to NDDs

Complementary works are aiming to face clinical and biological heterogeneity in NDDs using a subtyping approach based on imaging data. Using hierarchical clustering methods on neuroanatomical data, Hong and colleagues [95] identified three distinct morphometric subtypes in ASD: ASD-I characterized by cortical thickening, increased surface area, and tissue blurring; ASD-II with

cortical thinning and decreased geodesic distance; and ASD-III with increased geodesic distance. These groups were associated with gradual symptom severities and might help tackle the well-known clinical heterogeneity issue introduced in Subheading 1. The genetic contribution to the observed clinical heterogeneity was investigated across eight psychiatric conditions including ASD and ADHD [96] with common variants. Exploratory factor analysis (EFA) on GWAS cross-disorders' summary results led to the identification of three genetically inter-related groups of disorders, explaining together 51% of the genetic variation across NDDs and psychiatric conditions. The first factor linked anorexia nervosa, OCD, and Tourette syndrome. The second one was associated with major depression, bipolar disorder, and schizophrenia. The last one encompassed early-onset NDDs (ASD, ADHD, Tourette syndrome) and major depression. Similar to EFA results, hierarchical genetic clustering identified the same three subgroups among the eight disorders. These methods therefore have a great potential to uncover new biologically relevant diagnostic categories.

Such overlaps across clinical diagnoses have also been characterized at the imaging level. Patel et al. [19] determined a common pattern of group differences in cortical thickness across six disorders—including ASD, OCD, ADHD, schizophrenia, bipolar, and major depression disorders—and their link with gene expression profiles. Analyses of correlation and clustering revealed a shared profile of differences across disorders with 48% of variance explained, associated with pyramidal-cell gene expression. Analyses of gene co-expression highlighted two pre- and postnatal clusters associated with this common brain profile of group differences, enriched with genes associated with these disorders. Kebets and colleagues [97] applied partial least square regression (PLSR) to resting-state fMRI and cognitive metrics in participants with either ASD, ADHD, schizophrenia, or bipolar disorders. They identified three latent components (general psychopathology, cognitive dysfunction, and impulsivity) with unique fMRI signatures. Connectivity patterns of the somatosensory-motor network were main drivers across the three components. Similar findings on the somatosensory-motor network have been observed by [98] and extended to rare genetic mutations that confer high risk for neuropsychiatric conditions. Kernbach et al. [42] designed a hierarchical Bayesian modeling framework to derive hidden disease dimensions from RS-fMRI data across a population of ADHD, ASD, and controls. Using these methods, the number of components is inferred from the data. They obtained 45 hidden components that were then reduced to 3 main factors for better interpretation. For each of these three identified factors, the authors characterized the associated fMRI coupling patterns and symptom measures from the clinical questionnaires. These brain-

derived factors predicted the classification of subjects as ADHD, ASD, or control with an accuracy of 67%, computed using a variant of cross-validation called pre-validation described in [99]. This variant is expected to enable a fairer evaluation of the group labels than cross-validation, but still leaves room for errors compared to out-of-sample predictions [46].

Latent space decomposition techniques have been also used to identify general principles of the hierarchical brain organization—denoted as functional gradients—that locate sensory-motor networks at one end and the transmodal default-mode network at the other end [100, 101]. Hong and colleagues [102] hypothesized that NDDs' conditions may preferentially affect the sensory-motor dimension. They used surface-based analytical models to compare the first functional gradient (explaining 24% of the connectome variance) in ASD vs. controls and showed that both extremes of the rostrocaudal gradient were decreased in ASD. Interestingly, vertex-wise analyses revealed that such diminution in ASD was driven by transmodal medial PFC and posterior cingulate regions [102].

Combining large-scale multidimensional data is perceived as the golden standard to correctly apply ML algorithms. However, only a few precision medicine studies managed so far to do so. In [103], the authors extracted electronic health records, familial whole-exome sequences, and neurodevelopmental gene expression patterns in a large sample of ASD patients. Their goal was to identify biologically homogeneous ASD subtypes. For this purpose, the authors used spatiotemporal expression data from typically developing human brains to identify clusters of exons that are co-expressed during early human brain development. Based on prior knowledge on sexually different prenatal gene expression in ASD, they focused the analysis on a set of clusters that are differentially expressed between males and females. They then selected inherited, likely gene-disrupting variants among all the ASD-segregating ones by leveraging a large dataset of families who have one child with ASD and one unaffected sibling. They mapped variants back to exon clusters to identify 33 clusters of neurodevelopmentally co-regulated, ASD-segregating deleterious variants. The functional enrichment analysis of the identified exon clusters (detailed in [103]) revealed a new molecular convergence on lipid regulation, with variants expected to collectively alter LDL, cholesterol, and triglyceride levels. They confirmed that children with ASD have blood lipid profiles that are significantly outside the physiological range. Finally, they characterized the diagnostic spectrum of the dyslipidemia-associated ASD subtype and confirmed its specificity by comparing with individuals with ASD and no dyslipidemia. This work demonstrated the potential of combining massive amounts of multimodal data for uncovering new ASD subtypes.

3.3 Normative Modeling

In [104], the authors applied normative modeling to a large sample of ASD and controls males covering a wide age range (5–40 years). They investigated the potential of age-related effects on cortical thickness to serve as an individualized metric of atypicality in individuals with ASD. They reported that only a small subgroup of patients showed age-atypical cortical thickness. By comparing with conventional case-control analyses, they observed that most case-control differences were driven by a small subgroup of patients with high atypicality for their age. Highly consistent results were obtained in another application of normative modeling to a different ASD cohort [105], despite important variations across these studies. The population of the second work was composed of both males and females, and sex was included as a factor in the normative model. In addition, the normative models were estimated using different approaches (non-parametric regression in [104], Gaussian process regression in [105]). The overall consistent results despite the methodological differences support the relevance of the normative modeling approach for NDDs. In a follow-up study, [106] applied the spectral clustering technique to the atypicality maps computed at the individual level as deviation in the cortical thickness with respect to the normative model estimated in [105]. They identified five subtypes of individuals with ASD and assessed their separability using a multi-class linear SVM. Each subpopulation was then characterized in terms of demographic and clinical measures as well as association with polygenic scores for seven traits (autism, ADHD, epilepsy, full IQ, neuroticism, schizophrenia, and cross-disorder risk for psychiatric disorders). Importantly, they observed striking differences in the spatial patterns of cortical thickness atypicality maps between subtypes: three clusters showed reduced cortical thickness relative to the normative pattern, whereas two clusters showed an increased cortical thickness. These distinct and opposing atypicalities across different subtypes could explain the inconsistency in the previous case-control analyses. A last study did apply normative modeling to an adult population of ADHD patients [107]. The authors estimated a normative model predicting regional gray and white matter volumes across the brain from age and sex. They observed deviations shared across patients in gray matter in the cerebellum, temporal regions, and the hippocampus. They also provided a measure of the inter-individual variation between ADHD patients with extreme deviations in specific regions in more than 2% of the participants. Overall, these results highlighted the relevance of the normative modeling approach to understanding the heterogeneity in NDDs.

3.4 Genetic Features to Predict Cognitive Deficit in NDDs

As extensively discussed in [39], attempts to dissect mechanisms of NDD have mainly used a top-down approach, starting with a diagnosis and moving down to brain intermediate phenotypes and to genes. By contrast, the recruitment of groups based on the

presence of a genetic risk factor for NDDs allows for the investigation of pathways related to a particular biological risk for psychiatric symptoms (bottom-up approach). Clinical routine with genomic microarrays revealed that copy number variants are present in 10–15% of children with neurodevelopmental conditions [108]. Genetic-first approaches can however only be applied to a few recurrent pathogenic mutations frequent enough to establish a case-control study design. Thus, the effect of the vast majority of rare deleterious risk variants remains undocumented. Because a highly diverse landscape of rare variants confers a higher risk to a spectrum of NDDs, studies focusing on individual mutations will not be able to properly disentangle the relationship between mutations, molecular mechanisms, and diagnoses. Huguet and colleagues [109] speculated that large effect size pathogenic deletions may be attributable to the sum of individual effects of genes encompassed in each copy number variation. They introduced a new framework to estimate the effect of any pathogenic deletion on intelligence quotient (IQ). Using several types of functional annotations of rare genetic deletions associated with NDDs, the proposed framework predicted their impact on IQ with 76% accuracy [109]. They showed that haploinsufficiency scores—probability of being loss of function intolerant (pLI)—best explain the cognitive deficits. Follow-up works specifically on ASD confirmed that this score was the best predictor of IQ deficit and autism risk (odds ratio) [110, 111]. Deletion of 1 point of pLI was associated with a decrease of 2.6 points of IQ in autism.

3.5 Deep Learning Applied to NDDs

A deep learning-based framework has been recently introduced to predict the regulatory contribution of non-coding mutations to autism [112]. Authors constructed a deep convolutional network to model the functional impact of each individual mutation (single nucleotide polymorphism). They first identified that ASD probands ($n = 1700$ families) were carriers of a higher rate of transcriptional and post-transcriptional regulation disrupting de novo mutations compared with their siblings. They also revealed a convergent pattern of coding and non-coding mutations.

In [113], the authors analyzed resting-state fMRI (RS-fMRI) data from 260 subjects with ADHD and 343 healthy controls from the ADHD-200 database. They proposed to represent RS-fMRI data from each individual as a graph that integrates both temporal and spatial correlation of regional time-series signals. An original graph convolutional neural network architecture was introduced to characterize the brain functional connectome. The model also included seven non-imaging variables (age, gender, handedness, IQ measurement, and three Wechsler Intelligence Scale evaluation IQ variables) and was trained to distinguish ADHD patients from HC. Several experiments showed a performance gain compared to previous methods including SVM, logistic regression, and conventional graph convolutional networks. The proposed method

outperformed other competing approaches, including SVM and logistic regression, with an AUC of 75 (72.0% accuracy, 71.6% specificity, and 72.2% sensitivity) on a tenfold cross-validation. A leave-study-site-out experiment demonstrated the robustness of the proposed model for unseen data from different study sites, and experiments with simplified versions of the model showed the relevance of each proposed improvement. Most discriminative regions were mainly located in the frontal lobe, occipital lobe, subcortical lobe, temporal lobe, and cerebellum—with hypo-connections mainly between the frontal, parietal, and temporal lobes and widespread hyper-connections.

These studies support that new methodological improvements can be expected from the very active field of deep learning applications to neuroimaging and genetics data. As pointed in [88], the anticipated increase in sample size in NDDs studies will allow fit more complex models, which might reveal larger differences in performances compared to conventional methods. The literature of DL applications to NDDs is however still in its initial stages, and major challenges such as tendency to overfitting [43] have to be carefully addressed in future studies.

3.6 Discussion

The review of the selected recent studies presented above demonstrates that the application of ML in NDDs is a very active field of research, with encouraging perspectives. This field indeed benefits directly from initiatives to openly share data [114], which did increase the sample size involved across studies, and favored the engagement of ML scientists. The paradigm shift from diagnostic-first to genetic-first and from one diagnostic at a time to cross-diagnoses approaches is afoot, with a clear rise of large-scale studies based on normative modeling and deep learning approaches. Methodological works continue to introduce new innovative ML approaches specifically designed to address the central tasks in NDDs. Importantly, the adoption of best practices for the validation and replication of the results across independent datasets as stated in [46] is clearly encouraged by the recent reviews [31, 32, 34, 51, 55, 69, 71]. However, the validation is limited by insufficient access to large enough datasets combining multiscale data (genetics, transcriptomic, proteomic, metabolomic, neuroimaging features, phenomics). There is no open dataset so far offering that level of granularity. Indeed, the imaging field is just reaching the sample size allowing for running modern ML techniques for some but not all modalities. For instance, large-scale studies involving diffusion-weighted imaging are clearly lacking in NDDs, probably due to insufficient access to appropriate data. The genomic field is not ready yet, and several domains remain relatively new (e.g., first genome sequenced in 2000, next-generation sequencing techniques in 2010) and expensive (e.g., RNA-Seq data) [115, 116]. Such data will provide—in the near future—massive potential for accurate classification and appropriate validation.

4 Open Challenges and Conclusion

Methodological improvements described in Subheading 2 and studies reviewed in Subheading 3 are encouraging for concrete impact on clinical practice in the future. However, such clinical translation is raising major challenges that should be addressed.

4.1 Potential Bias in Data and Processing Pipelines

Despite the large amount of new approaches released by recent literature, some potential biases in analysis pipelines should be mentioned. For instance, the analysis of functional networks computed from RS-fMRI relies on a complex succession of processing steps. Several of these processing steps actually correspond to implementing assumptions regarding the data. However, the validity of these assumptions and their influence on the subsequent results are not sufficiently discussed in the literature. See, for instance, [117] for a quantitative evaluation of the impact of the brain parcellation procedure on functional connectivity analyses. Another major barrier to reproducibility is the lack of compatibility among programming languages, software versions, and operating systems as illustrated in [118]. This report highlights the challenges and potential solutions to be implemented at both the individual researcher and community levels in order to enable the appropriate reuse of published methods.

On the data side, the limitations related to the absence of recording of potentially influencing factors are not sufficiently investigated and acknowledged. As pointed, e.g., in [119]: “The extent of brain differences in disease may depend critically on a patient’s age, duration of illness, course of treatment, as well as adherence to the treatment, polypharmacy and other unmeasured factors. Differences in ancestral background, as determined based on genotype, are strongly related to systematic differences in brain shape. Any realistic understanding of the brain imaging measures must take all these into account, as well as acknowledge the existence of causal factors perhaps not yet known or even imagined.” As a concrete illustration, [120, 121] recently reported significant alterations in brain morphometry induced by prematurity, a factor that was not considered by any of the studies we reviewed here. Such uncontrolled factors might introduce considerable bias in the learning process. The ML research field has identified this pitfall, and several solutions to prevent unexpected implications in clinical applications are actively debated [122–124].

4.2 Interpretability and Biological Substrates

Even in the absence of bias, the interpretation of the outcome of any ML algorithm in the context of clinical application represents a critical challenge. More than the level of raw performance, the level of expertise required from medical doctors in (1) the recording and (2) the analysis of the data compared to “expertise-free” raw data is

a question that requires more attention. We refer to [125] for a thoughtful discussion on the need for clarification of the role of ML-based tools in relation to clinicians' decisions and actions in clinical practice. The authors call for a more systematic demonstration that models learning from non-clinician-initiated data outperform models based on clinician-initiated data. They purposely argue that models driven by features derived from the actions of clinicians and not related to the underlying physiology might introduce some deleterious circularity. Indeed, the outcome of such a model might potentially confuse more than support a clinician in his decisions.

Then—regarding the interpretation in terms of pathophysiology—the challenge is to relate the decisions of any ML techniques to putative underlying biological processes. Methodological innovations will enhance the explainability of ML models, but explainability and transparency do not imply interpretability [126, 127]. Another major challenge is to assess the biological relevance of the features extracted from the data during the learning procedure. Purely data-driven approaches are limited by the difficulty to relate the parameters of the model to biological knowledge. A promising perspective consists in inserting biological priors directly in predictive models. See [92] for an introductory review to this type of approach in the context of computational neuroscience and (<https://sciml.ai>) for further information on the emerging field of scientific machine learning. However, extensive basic research at conceptual, methodological, and experimental levels are required to fill the gap between measures accessible *in vivo* in patients and the biophysiology acting at cellular and molecular levels. See, for instance, [128] for an illustration of the complexity of this challenge, where the authors propose a framework integrating different levels of interactions, from genes to cells, circuits, and clinical expression, to better understand and treat cortical malformations. As discussed in [129] for ASD, research designs aiming at a better conceptual integration between different levels of brain organization are required to characterize the cascade of pathogenic processes in NDDs.

4.3 Conclusion

In NDDs, as in healthcare in general, ML has a role to play in addressing the longstanding deficiencies such as serious diagnostic errors, mistakes in treatment, and waste of resources [130]. Indeed, ML will undoubtedly help redefine NDDs' categories and other mental illnesses more objectively, identify them at an early stage, and contribute to more adapted treatments. The rise of ML is the occasion to improve the standardization of practice and to enforce the generalization of open science with preregistration and data sharing or federated learning. In addition, the field has to demonstrate high and reproducible performances in the real-world clinical environment. Finally, major conceptual, ethical, and socio-technical challenges have to be addressed.

Acknowledgments

We would like to thank the editor and Guillaume Dumas who served as a reviewer for their insightful feedback that has improved our manuscript. This work was supported by the French government under the management of Agence Nationale de la Recherche, reference ANR-19-CE45-0014.

References

1. American Psychiatric Association (2013) The diagnostic and statistical manual of mental disorders: DSM 5. American Psychiatric Publishing, Arlington, VA
2. Jacob S, Wolff JJ, Steinbach MS, Doyle CB, Kumar V, Elison JT (2019) Neurodevelopmental heterogeneity and computational approaches for understanding autism. *Transl Psychiatry* 9(1):1–12. <https://doi.org/10.1038/s41398-019-0390-0>
3. Lombardo MV, Lai M-C, Baron-Cohen S (2019) Big data approaches to decomposing heterogeneity across the autism spectrum. *Mol Psychiatry* 24(10):1435. <https://doi.org/10.1038/s41380-018-0321-0>
4. Hyman SE (2007) Can neuroscience be integrated into the DSM-V? *Nat Rev Neurosci* 8(9):725–732. <https://doi.org/10.1038/nrn2218>
5. Bourgeron T (2015) What do we know about early onset neurodevelopmental disorders? <https://doi.org/10.7551/mitpress/9780262029865.003.0005>
6. Joshi G et al (2010) The heavy burden of psychiatric comorbidity in youth with autism spectrum disorders: a large comparative study of a psychiatrically referred population. *J Autism Dev Disord* 40(11):1361–1370. <https://doi.org/10.1007/s10803-010-0996-9>
7. Lai M-C, Lombardo MV, Baron-Cohen S (2014) Autism. *Lancet* 383(9920):896–910. [https://doi.org/10.1016/S0140-6736\(13\)61539-1](https://doi.org/10.1016/S0140-6736(13)61539-1)
8. Anttila V et al (2018) Analysis of shared heritability in common disorders of the brain. *Science* 360(6395):eaap8757. <https://doi.org/10.1126/science.aap8757>
9. Siugzdaite R, Bathelt J, Holmes J, Astle DE (2020) Transdiagnostic brain mapping in developmental disorders. *Curr Biol* 30(7):1245–1257.e4. <https://doi.org/10.1016/j.cub.2020.01.078>
10. Leblond CS et al (2021) Operative list of genes associated with autism and neurodevelopmental disorders based on database review. *Mol Cell Neurosci* 113:103623. <https://doi.org/10.1016/j.mcn.2021.103623>
11. Mills KL et al (2021) Inter-individual variability in structural brain development from late childhood to young adulthood. *NeuroImage* 242:118450. <https://doi.org/10.1016/j.neuroimage.2021.118450>
12. Brown TT (2017) Individual differences in human brain development. *Wiley Interdiscip Rev Cogn Sci* 8(1–2):e1389. <https://doi.org/10.1002/wcs.1389>
13. Brown TT et al (2012) Neuroanatomical assessment of biological maturity. *Curr Biol* 22(18):1693–1698. <https://doi.org/10.1016/j.cub.2012.07.002>
14. Thompson DK et al (2020) Tracking regional brain growth up to age 13 in children born term and very preterm. *Nat Commun* 11(1):696. <https://doi.org/10.1038/s41467-020-14334-9>
15. Witvliet D et al (2021) Connectomes across development reveal principles of brain maturation. *Nature* 596(7871):257–261. <https://doi.org/10.1038/s41586-021-03778-8>
16. Fjell AM et al (2019) Continuity and discontinuity in human cortical development and change from embryonic stages to old age. *Cereb Cortex* 29(9):3879–3890. <https://doi.org/10.1093/cercor/bhy266>
17. Reh RK et al (2020) Critical period regulation across multiple timescales. *Proc Natl Acad Sci* 117(38):23242–23251. <https://doi.org/10.1073/pnas.1820836117>
18. Rudel RG (1981) Residual effects of childhood reading disabilities. *Bull Orton Soc* 31:89–102
19. Patel Y et al (2020) Virtual histology of cortical thickness and shared neurobiology in 6 psychiatric disorders. *JAMA Psychiatry*. <https://doi.org/10.1001/jamapsychiatry.2020.2694>
20. Casey BJ et al (2018) The Adolescent Brain Cognitive Development (ABCD) study:

- imaging acquisition across 21 sites. *Dev Cogn Neurosci* 32(January):43–54. <https://doi.org/10.1016/j.dcn.2018.03.001>
21. Di Martino A et al (2014) The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol Psychiatry* 19(April):659–667. <https://doi.org/10.1038/mp.2013.78>
 22. Loth E et al (2017) The EU-AIMS longitudinal European Autism Project (LEAP): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders. *Mol Autism* 8(1):24. <https://doi.org/10.1186/s13229-017-0146-8>
 23. Bellec P, Chu C, Chouinard-Decorte F, Benhajali Y, Margulies DS, Craddock RC (2017) The neuro bureau ADHD-200 pre-processed repository. *NeuroImage* 144:275–286. <https://doi.org/10.1016/j.neuroimage.2016.06.034>
 24. May T et al (2018) Trends in the overlap of autism spectrum disorder and attention deficit hyperactivity disorder: prevalence, clinical management, language and genetics. *Curr Dev Disord Rep* 5(1):49–57. <https://doi.org/10.1007/s40474-018-0131-8>
 25. Mansour R, Dovi AT, Lane DM, Loveland KA, Pearson DA (2017) ADHD severity as it relates to comorbid psychiatric symptomatology in children with autism spectrum disorders (ASD). *Res Dev Disabil* 60:52–64. <https://doi.org/10.1016/j.ridd.2016.11.009>
 26. Hoogman M et al (2022) Consortium neuroscience of attention deficit/hyperactivity disorder and autism spectrum disorder: the ENIGMA adventure. *Hum Brain Mapp* 43(1):37. <https://doi.org/10.1002/hbm.25029>
 27. Fombonne E (1999) The epidemiology of autism: a review. *Psychol Med* 29(4): 769–786. <https://doi.org/10.1017/S0033291799008508>
 28. Bourgeron T (2015) From the genetic architecture to synaptic plasticity in autism spectrum disorder. *Nat Rev Neurosci* 16(9): 551–563. <https://doi.org/10.1038/nrn3992>
 29. Lord C et al (2020) Autism spectrum disorder. *Nat Rev Dis Primer* 6(1). <https://doi.org/10.1038/s41572-019-0138-4>
 30. Thapar A, Cooper M (2016) Attention deficit hyperactivity disorder. *Lancet* 387(10024): 1240–1250. [https://doi.org/10.1016/S0140-6736\(15\)00238-X](https://doi.org/10.1016/S0140-6736(15)00238-X)
 31. Wolfers T et al (2019) From pattern classification to stratification: towards conceptualizing the heterogeneity of autism spectrum disorder. *Neurosci Biobehav Rev* 104 (April):240–254. <https://doi.org/10.1016/j.neubiorev.2019.07.010>
 32. Xu M, Calhoun V, Jiang R, Yan W, Sui J (2021) Brain imaging-based machine learning in autism spectrum disorder: methods and applications. *J Neurosci Methods* 361: 109271. <https://doi.org/10.1016/j.jneumeth.2021.109271>
 33. Hiremath CS et al (2021) Emerging behavioral and neuroimaging biomarkers for early and accurate characterization of autism spectrum disorders: a systematic review. *Transl Psychiatry* 11(1):1–12. <https://doi.org/10.1038/s41398-020-01178-6>
 34. Bzdok D, Meyer-Lindenberg A (2018) Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging* 3(3): 223–230. <https://doi.org/10.1016/j.bpsc.2017.11.007>
 35. Hyde KK et al (2019) Applications of supervised machine learning in autism spectrum disorder research: a review. *Rev J Autism Dev Disord* 6(2):128–146. <https://doi.org/10.1007/s40489-019-00158-x>
 36. Eslami T, Almuqhim F, Raiker JS, Saeed F (2021) Machine learning methods for diagnosing autism spectrum disorder and attention- deficit/hyperactivity disorder using functional and structural MRI: a survey. *Front Neuroinform* 14. Accessed 21 Jan 2022. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fninf.2020.575999>
 37. Mottron L, Bzdok D (2020) Autism spectrum heterogeneity: fact or artifact? *Mol Psychiatry* 25(12):3178–3185. <https://doi.org/10.1038/s41380-020-0748-y>
 38. Loth E et al (2021) The meaning of significant mean group differences for biomarker discovery. *PLoS Comput Biol* 17(11): e1009477. <https://doi.org/10.1371/journal.pcbi.1009477>
 39. Moreau CA, Raznahan A, Bellec P, Chakravarty M, Thompson PM, Jacquemont S (2021) Dissecting autism and schizophrenia through neuroimaging genomics. *Brain* 144: 1943. <https://doi.org/10.1093/brain/awab096>
 40. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22(11):1359–1366. <https://doi.org/10.1177/0956797611417632>

41. Marquand AF, Kia SM, Zabihi M, Wolfers T, Buitelaar JK, Beckmann CF (2019) Conceptualizing mental disorders as deviations from normative functioning. *Mol Psychiatry* 24: 1415. <https://doi.org/10.1038/s41380-019-0441-1>
42. Kernbach JM et al (2018) Shared endophenotypes of default mode dysfunction in attention deficit/hyperactivity disorder and autism spectrum disorder. *Transl Psychiatry* 8(1):1–11. <https://doi.org/10.1038/s41398-018-0179-6>
43. Traut N et al (2022) Insights from an autism imaging biomarker challenge: promises and threats to biomarker discovery. *NeuroImage* 255:119171. <https://doi.org/10.1016/j.neuroimage.2022.119171>
44. Maier-Hein L et al (2018) Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat Commun* 9(1). <https://doi.org/10.1038/s41467-018-07619-7>
45. Bron EE et al (2021) Ten years of image analysis and machine learning competitions in dementia. *ArXiv211207922* Cs. Accessed 21 Dec 2021. [Online]. Available: <http://arxiv.org/abs/2112.07922>
46. Poldrack RA, Huckins G, Varoquaux G (2020) Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* 77(5):534–540. <https://doi.org/10.1001/jamapsychiatry.2019.3671>
47. Varoquaux G (2018) Cross-validation failure: small sample sizes lead to large error bars. *NeuroImage* 180:68–77. <https://doi.org/10.1016/j.neuroimage.2017.06.061>
48. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B (2017) Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage* 145:166–179. <https://doi.org/10.1016/j.neuroimage.2016.10.038>
49. Bouthillier X et al (2021) Accounting for variance in machine learning benchmarks. *ArXiv210303098* Cs Stat. Accessed 21 Dec 2021. [Online]. Available: <http://arxiv.org/abs/2103.03098>
50. Kassraian-Fard P, Matthis C, Balsters JH, Maathuis MH, Wenderoth N (2016) Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. *Front Psych* 7:177. <https://doi.org/10.3389/fpsyg.2016.00177>
51. Bzdok D, Ioannidis JPA (2019) Exploration, inference, and prediction in neuroscience and biomedicine. *Trends Neurosci* 42(4): 251–262. <https://doi.org/10.1016/j.tins.2019.02.001>
52. Abraham A et al (2017) Deriving reproducible biomarkers from multi-site resting-state data: an Autism-based example. *NeuroImage* 147(October 2016):736–745. <https://doi.org/10.1016/j.neuroimage.2016.10.045>
53. Johansen-Berg H, Behrens TEJ (2014) Diffusion MRI. Academic p. Elsevier. <https://doi.org/10.1016/C2011-0-07047-3>
54. Choi SW, Mak TS-H, O'Reilly PF (2020) Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* 15(9): 2759–2772. <https://doi.org/10.1038/s41596-020-0353-1>
55. Uddin LQ, Dajani DR, Voorhies W, Bednarz H, Kana RK (2017) Progress and roadblocks in the search for brain-based biomarkers of autism and attention-deficit/hyperactivity disorder. *Transl Psychiatry* 7(8). <https://doi.org/10.1038/tp.2017.164>
56. Cameron C et al (2013) The neuro bureau preprocessing initiative: open sharing of pre-processed neuroimaging data and derivatives. *Front Neuroinform* 7. <https://doi.org/10.3389/conf.fniinf.2013.09.00041>
57. Rødgaard E-M, Jensen K, Vergnes J-N, Soulières I, Mottron L (2019) Temporal changes in effect sizes of studies comparing individuals with and without autism: a meta-analysis. *JAMA Psychiatry* 76(11): 1124–1132. <https://doi.org/10.1001/jamapsychiatry.2019.1956>
58. Insel T et al (2010) Research Domain Criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry* 167(7):748–751. <https://doi.org/10.1176/appi.ajp.2010.09091379>
59. Insel TR, Cuthbert BN (2015) Brain disorders? Precisely. *Science* 348(6234):499–500. <https://doi.org/10.1126/science.aab2358>
60. Cuthbert BN, Insel TR (2013) Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med* 11(1):126. <https://doi.org/10.1186/1741-7015-11-126>
61. Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E (2001) The Autism-Spectrum Quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J Autism Dev Disord* 31(1): 5–17. <https://doi.org/10.1007/s10802-001-0001-7>
62. Constantino JN, Gruber CP (2012) Social responsiveness scale: SRS-2. Western Psychological Services, Torrance, CA

63. Ronald A, Hoekstra RA (2011) Autism spectrum disorders and autistic traits: a decade of new twin studies. *Am J Med Genet B Neuropsychiatr Genet* 156(3):255–274. <https://doi.org/10.1002/ajmg.b.31159>
64. Feczko E, Miranda-Dominguez O, Marr M, Graham AM, Nigg JT, Fair DA (2019) The heterogeneity problem: approaches to identify psychiatric subtypes. *Trends Cogn Sci* 23(7):584–601. <https://doi.org/10.1016/j.tics.2019.03.009>
65. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7_1
66. Tang S, Sun N, Floris DL, Zhang X, Martino AD, Yeo BTT (2020) Reconciling dimensional and categorical models of autism heterogeneity: a brain connectomics and behavioral study. *Biol Psychiatry* 87(12):1071–1082. <https://doi.org/10.1016/j.biopsych.2019.11.009>
67. Feczko E, Fair DA (2020) Methods and challenges for assessing heterogeneity. *Biol Psychiatry* 88(9). <https://doi.org/10.1016/j.biopsych.2020.02.015>
68. von Luxburg U, Williamson RC, Guyon I (2012) Clustering: science or art? in Proceedings of ICML Workshop on Unsupervised and Transfer Learning, pp. 65–79. Accessed 12 Jan 2022. [Online]. Available: <https://proceedings.mlr.press/v27/luxburg12a.html>
69. Hong S-J et al (2020) Toward neurosubtypes in autism. *Biol Psychiatry* 88(1):111–128. <https://doi.org/10.1016/j.biopsych.2020.03.022>
70. Nordahl CW et al (2022) The autism phenotype project: toward identifying clinically meaningful subgroups of autism. *Front Neurosci* 15. Accessed 21 Jan 2022. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fnins.2021.786220>
71. Agelink van Rentergem JA, Deserno MK, Geurts HM (2021) Validation strategies for subtypes in psychiatry: a systematic review of research on autism spectrum disorder. *Clin Psychol Rev* 87:102033. <https://doi.org/10.1016/j.cpr.2021.102033>
72. Bzdok D, Yeo BTT (2017) Inference in the age of big data: future perspectives on neuroscience. *NeuroImage* 155(April):549–564. <https://doi.org/10.1016/j.neuroimage.2017.04.061>
73. Grotzinger AD et al (2019) Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat Hum Behav* 3(5):513. <https://doi.org/10.1038/s41562-019-0566-x>
74. Modenato C et al (2021) Lessons learnt from neuroimaging studies of Copy Number Variants, a systematic review. *Biol Psychiatry* 90: S0006322321013949. <https://doi.org/10.1016/j.biopsych.2021.05.028>
75. Marquand AF, Rezek I, Buitelaar J, Beckmann CF (2016) Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biol Psychiatry* 80(7):552–561. <https://doi.org/10.1016/j.biopsych.2015.12.023>
76. Williams CK, Rasmussen CE (2006) Gaussian processes for machine learning, vol 2. MIT Press, Cambridge, MA
77. Xu B, Kuplicki R, Sen S, Paulus MP (2021) The pitfalls of using Gaussian Process Regression for normative modeling. *PLoS One* 16(9):e0252108. <https://doi.org/10.1371/journal.pone.0252108>
78. Lefebvre A et al (2018) Alpha waves as a biomarker of autism spectrum disorder: the challenge of reproducibility and heterogeneity. *Front Neurosci* 12. Accessed 10 May 2022. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fnins.2018.00662>
79. Chen G et al (2021) Beyond linearity in neuroimaging: capturing nonlinear relationships with application to longitudinal studies. *NeuroImage* 233:117891. <https://doi.org/10.1016/j.neuroimage.2021.117891>
80. Fraza CJ, Dinga R, Beckmann CF, Marquand AF (2021) Warped Bayesian linear regression for normative modelling of big data. *NeuroImage* 245:118715. <https://doi.org/10.1016/j.neuroimage.2021.118715>
81. Fjell AM et al (2010) When does brain aging accelerate? Dangers of quadratic fits in cross-sectional studies. *NeuroImage* 50(4):1376–1383. <https://doi.org/10.1016/j.neuroimage.2010.01.061>
82. Fjell AM et al (2015) Development and aging of cortical thickness correspond to genetic organization patterns. *Proc Natl Acad Sci* 112(50):15462–15467. <https://doi.org/10.1073/pnas.1508831112>
83. Dinga R, Fraza CJ, Bayer JMM, Kia SM, Beckmann CF, Marquand AF (2021) Normative modeling of neuroimaging data using generalized additive models of location scale and shape. *bioRxiv*, p. 2021.06.14.448106. <https://doi.org/10.1101/2021.06.14.448106>
84. Bethlehem RAI et al (2022) Brain charts for the human lifespan. *Nature* 604(7906):525.

- <https://doi.org/10.1038/s41586-022-04554-y>
85. Bataille D, Edwards AD, O'Muircheartaigh J (2018) Annual research review: not just a small adult brain: understanding later neurodevelopment through imaging the neonatal brain. *J Child Psychol Psychiatry* 59(4): 350–371. <https://doi.org/10.1111/jcpp.12838>
 86. Rutherford S et al (2021) The normative modeling framework for computational psychiatry. *bioRxiv*, p. 2021.08.08.455583. <https://doi.org/10.1101/2021.08.08.455583>
 87. Rutherford S et al (2022) Charting brain growth and aging at high spatial precision. *eLife* 11:e72904. <https://doi.org/10.7554/eLife.72904>
 88. Vieira S, Pinaya WHL, Mechelli A (2017) Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci Biobehav Rev* 74:58–75. <https://doi.org/10.1016/j.neubiorev.2017.01.002>
 89. Abrol A et al (2021) Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat Commun* 12(1). <https://doi.org/10.1038/s41467-020-20655-6>
 90. Zabihí M et al (2021) Non-linearity matters: a deep learning solution to generalization of hidden brain patterns across population cohorts. <https://doi.org/10.1101/2021.03.10.434856>
 91. Pinaya WHL, Mechelli A, Sato JR (2019) Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: a large-scale multi-sample study. *Hum Brain Mapp* 40(3):944–954. <https://doi.org/10.1002/hbm.24423>
 92. Panahi MR, Abrevaya G, Gagnon-Audet J-C, Voleti V, Rish I, Dumas G (2021) Generative models of brain dynamics – a review. *ArXiv211212147 Q-Bio*. Accessed 10 May 2022. [Online]. Available: <http://arxiv.org/abs/2112.12147>
 93. Abrevaya G et al (2021) Learning brain dynamics with coupled low-dimensional non-linear oscillators and deep recurrent networks. *Neural Comput* 33(8):2087–2127. https://doi.org/10.1162/neco_a_01401
 94. Hazlett HC et al (2017) Early brain development in infants at high risk for autism spectrum disorder. *Nature* 542(7641):348–351. <https://doi.org/10.1038/nature21369>
 95. Hong S-J, Valk SL, Di Martino A, Milham MP, Bernhardt BC (2017) Multidimensional neuroanatomical subtyping of autism spectrum disorder. *Cereb Cortex*, no. Betancur 2011, pp. 1–11. <https://doi.org/10.1093/cercor/bhx229>
 96. Lee PH et al (2019) Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* 179(7): 1469–1482.e11. <https://doi.org/10.1016/j.cell.2019.11.020>
 97. Kebets V et al (2019) Somatosensory-motor dysconnectivity spans multiple transdiagnostic dimensions of psychopathology. *Biol Psychiatry* 86(10):779–791. <https://doi.org/10.1016/j.biopsych.2019.06.013>
 98. Moreau CA et al (2020) Mutations associated with neuropsychiatric conditions delineate functional brain connectivity dimensions contributing to autism and schizophrenia. *Nat Commun* 11(1). <https://doi.org/10.1038/s41467-020-18997-2>
 99. Tibshirani RJ, Efron B (2002) Pre-validation and inference in microarrays. *Stat Appl Genet Mol Biol* 1(1). <https://doi.org/10.2202/1544-6115.1000>
 100. Huntenburg JM, Bazin P-L, Margulies DS (2018) Large-scale gradients in human cortical organization. *Trends Cogn Sci* 22(1): 21–31. <https://doi.org/10.1016/j.tics.2017.11.002>
 101. Margulies DS et al (2016) Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc Natl Acad Sci* 113(44):12574–12579. <https://doi.org/10.1073/pnas.1608282113>
 102. Hong S-J et al (2019) Atypical functional connectome hierarchy in autism. *Nat Commun* 10(1). <https://doi.org/10.1038/s41467-019-08944-1>
 103. Luo Y et al (2020) A multidimensional precision medicine approach identifies an autism subtype characterized by dyslipidemia. *Nat Med* 26(1):1375. <https://doi.org/10.1038/s41591-020-1007-0>
 104. Bethlehem RAI, Seidlitz J, Romero-Garcia R, Trakoshis S, Dumas G, Lombardo MV (2020) A normative modelling approach reveals age-atypical cortical thickness in a subgroup of males with autism spectrum disorder. *Commun Biol* 3(1):486. <https://doi.org/10.1038/s42003-020-01212-9>
 105. Zabihí M et al (2019) Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. *Biol Psychiatry Cogn Neurosci Neuroimaging* 4(6): 567–578. <https://doi.org/10.1016/j.bpsc.2018.11.013>

106. Zabihí M et al (2020) Fractionating autism based on neuroanatomical normative modeling. *Transl Psychiatry* 10(1):384. <https://doi.org/10.1038/s41398-020-01057-0>
107. Wolfers T, Beckmann CF, Hoogman M, Buitelaar JK, Franke B, Marquand AF (2020) Individual differences v. the average patient: mapping the heterogeneity in ADHD using normative models. *Psychol Med* 50(2): 314–323. <https://doi.org/10.1017/S0033291719000084>
108. Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST (2011) American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet Med* 13(7):680–685. <https://doi.org/10.1097/GIM.0b013e3182217a3a>
109. Huguet G et al (2018) Measuring and estimating the effect sizes of copy number variants on general intelligence in community-based samples. *JAMA Psychiatry* 75(5): 447–457. <https://doi.org/10.1001/jamapsychiatry.2018.0039>
110. Douard E et al (2020) Effect sizes of deletions and duplications on autism risk across the genome. *Am J Psychiatry* 178(1):87–98. <https://doi.org/10.1176/appi.ajp.2020.19080834>
111. Huguet G et al (2021) Genome-wide analysis of gene dosage in 24,092 individuals estimates that 10,000 genes modulate cognitive ability. *Mol Psychiatry* 26(6):2663. <https://doi.org/10.1038/s41380-020-00985-z>
112. Zhou J et al (2019) Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat Genet* 51(6):973. <https://doi.org/10.1038/s41588-019-0420-0>
113. Zhao K, Duka B, Xie H, Oathes DJ, Calhoun V, Zhang Y (2022) A dynamic graph convolutional neural network framework reveals new insights into connectome dysfunctions in ADHD. *NeuroImage* 246: 118774. <https://doi.org/10.1016/j.neuroimage.2021.118774>
114. Milham MP et al (2018) Assessment of the impact of shared brain imaging data on the scientific literature. *Nat Commun* 9(1):2818. <https://doi.org/10.1038/s41467-018-04976-1>
115. Tărlungeanu DC, Novarino G (2018) Genomics in neurodevelopmental disorders: an avenue to personalized medicine. *Exp Mol Med* 50(8):1. <https://doi.org/10.1038/s12276-018-0129-7>
116. Dias CM, Walsh CA (2020) Recent advances in understanding the genetic architecture of autism. *Annu Rev Genomics Hum Genet* 21(1):289–304. <https://doi.org/10.1146/annurev-genom-121219-082309>
117. Bryce NV et al (2021) Brain parcellation selection: an overlooked decision point with meaningful effects on individual differences in resting-state functional connectivity. *NeuroImage* 243:118487. <https://doi.org/10.1016/j.neuroimage.2021.118487>
118. Kim Y-M, Poline J-B, Dumas G (2018) Experimenting with reproducibility: a case study of robustness in bioinformatics. *GigaScience* 7(7):giy077. <https://doi.org/10.1093/gigascience/giy077>
119. Thompson PM et al (2017) ENIGMA and the individual: predicting factors that affect the brain in 35 countries worldwide. *NeuroImage* 145:389–408. <https://doi.org/10.1016/j.neuroimage.2015.11.057>
120. Dimitrova R et al (2020) Phenotyping the preterm brain: characterising individual deviations from normative volumetric development in two large infant cohorts. *Neuroscience*, preprint, <https://doi.org/10.1101/2020.08.05.228700>
121. Dimitrova R et al (2021) Preterm birth alters the development of cortical microstructure and morphology at term-equivalent age. *NeuroImage* 243:118488. <https://doi.org/10.1016/j.neuroimage.2021.118488>
122. Caton S, Haas C (2020) Fairness in machine learning: a survey. ArXiv201004053 Cs Stat. Accessed 07 Mar 2022. [Online]. Available: <http://arxiv.org/abs/2010.04053>
123. Mhasawade V, Zhao Y, Chunara R (2021) Machine learning and algorithmic fairness in public and population health. *Nat Mach Intell* 3(8):659. <https://doi.org/10.1038/s42256-021-00373-4>
124. Lee EE et al (2021) Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biol Psychiatry Cogn Neurosci Neuroimaging* 6(9):856–864. <https://doi.org/10.1016/j.bpsc.2021.02.001>
125. Beaulieu-Jones BK et al (2021) Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *Npj Digit Med* 4(1):1–6. <https://doi.org/10.1038/s41746-021-00426-3>
126. Boscolo Galazzo I et al (2022) Explainable artificial intelligence for magnetic resonance imaging aging brainprints: grounds and challenges. *IEEE Signal Process Mag* 39(2):

- 99–116. <https://doi.org/10.1109/MSP.2021.3126573>
127. Ghassemi M, Oakden-Rayner L, Beam AL (2021) The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 3(11):e745–e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
128. Klingler E, Francis F, Jabaudon D, Cappello S (2021) Mapping the molecular and cellular complexity of cortical malformations. *Science* 371(6527). <https://doi.org/10.1126/science.aba4517>
129. Courchesne E, Prampano T, Gazestani VH, Lombardo MV, Pierce K, Lewis NE (2019) The ASD living biology: from cell proliferation to clinical phenotype. *Mol Psychiatry* 24(1):88–107. <https://doi.org/10.1038/s41380-018-0056-y>
130. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 25(1):44. <https://doi.org/10.1038/s41591-018-0300-7>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Chapter 32

Machine Learning and Brain Imaging for Psychiatric Disorders: New Perspectives

Ivan Brossollet, Quentin Gallet, Pauline Favre, and Josselin Houenou

Abstract

Psychiatric disorders include a broad panel of heterogeneous conditions. Among the most severe psychiatric diseases, in intensity and incidence, depression will affect 15–20% of the population in their lifetime, schizophrenia 0.7–1%, and bipolar disorder 1–2.5%. Today, the diagnosis is solely based on clinical evaluation, causing major issues since it is subjective and as different diseases can present similar symptoms. These limitations in diagnosis lead to limitations in the classification of psychiatric diseases and treatments. There is therefore a great need for new biomarkers, usable at an individual level. Among them, magnetic resonance imaging (MRI) allows to measure potential brain abnormalities in patients with psychiatric disorders. This creates datasets with high dimensionality and very subtle variations between healthy subjects and patients, making machine and statistical learning ideal tools to extract biomarkers from these data. Machine learning brings different tools that could be useful to tackle these issues. On the one hand, supervised learning can support automated classification between different psychiatric conditions. On the other hand, unsupervised learning could allow the identification of new homogeneous subgroups of patients, refining our understanding of the classification of these disorders. In this chapter, we will review current research applying machine learning tools to brain imaging in psychiatry, and we will discuss its interest, limitations, and future applications.

Key words Psychiatry, Depression, Schizophrenia, Bipolar disorder, Machine learning, Artificial intelligence, Neuroimaging, MRI, Clustering, Classification

1 Introduction

Major psychiatric conditions affecting adults can be classified into several groups: affective disorders (e.g., bipolar disorders, major depressive disorders), psychotic disorders (e.g., schizophrenia), anxiety disorders (e.g., obsessive-compulsive disorders), neurodevelopmental disorders (e.g., autism), and substance use disorders. We will focus this chapter on the two first categories, as they carry a high individual and societal burden and are highly prevalent throughout the world.

1.1 Major Depressive Disorder

Major depressive disorder (MDD) is defined by the occurrence of one or more major depressive episodes without any manic or hypomanic episodes in the lifetime. Its prevalence can vary significantly according to the studies, but exceeds 15% of the population during their lifetime [1], and affects two women for one man. Depression can affect people at any time during their life [2]. Nowadays, the diagnosis is based on structured interviews, and the clinical criteria are given by, among others, two classification manuals: the International Classification of Diseases [1] and the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [3]. According to the DSM-5, to meet the criteria for a major depressive episode, five of the nine following symptoms must be present over a 2-week period: depressed mood or anhedonia (loss of interest or pleasure), change in weight or appetite, sleep disturbances (insomnia or hypersomnia), psychomotor retardation or restlessness, loss of energy or fatigue, low self-esteem or guilt, difficulty in concentrating or indecisiveness, and thoughts of death or suicidal thoughts. Patients with MDD are at an increased risk of other comorbid disorders. Most commonly, they may present alcohol abuse or dependence, anxiety disorders such as panic disorder, obsessive-compulsive disorder, and generalized anxiety disorder. Treatment options for MDD include a variable combination of pharmacotherapy (antidepressants such as serotonin selective reuptake inhibitors or tricyclics) and psychotherapy (cognitive behavioral therapy, interpersonal therapy, etc.). Despite considerable progress in its diagnosis and treatment, MDD remains underdiagnosed and underestimated and remains a challenge for healthcare institutions, especially since one of the main risks of mood disorders (BD or MDD) is suicidal behavior.

1.2 Bipolar Disorder

Bipolar disorder (BD) is defined as a chronic mood disorder characterized by episodes of depression and episodes of abnormal excitation (mania, hypomania), separated by periods of “euthymia” (without any symptoms of major mood episode) [3]. This mood disorder affects around 1% of the world’s adult population [4], regardless of continent, socioeconomic status, or ethnicity. The course of BD is lifelong, but is heterogeneous in terms of number of episodes, relapses, polarity (i.e., higher number of manic or depressive episodes), and response to treatment. The impact of the disease on cognitive function and quality of life can be major [4]. Diagnosis, treatment, health, and social care are major goals in the management of BD.

Manic episodes are defined by a period lasting at least 1 week, during which patients exhibit elevated mood and increased motor activity. The intensity of these symptoms defines the manic or hypomanic nature of the episode. During a manic episode, patients may experience psychotic symptoms such as hallucinations, delusions, disorganized thinking, and sleep disturbances. The delusions may be consistent with the manic mood, with individuals displaying

grandiosity, megalomania, or messianic ideas. Impaired judgment and risk of endangering the patient often lead to hospitalization. Hypomanic episodes are characterized by lower symptom intensity (abnormally high, expansive, or irritable mood, as well as abnormal increase in activity or energy, most of the day) and must last at least 4 consecutive days. Although there are no pathognomonic features of bipolar or unipolar depression, some clinical features are useful in distinguishing them: bipolar depression usually occurs at an earlier age, and the episodes are also more frequent and shorter, show an abrupt onset and termination, and are more frequently associated with substance abuse. Patients with bipolar depression may also present atypical symptoms, such as hypersomnia and weight instability. Psychosis (delusions and hallucinations) and catatonia are also more frequent in bipolar depression, whereas somatic complaints are more common in unipolar depression. The presence of a family history of mania is also a relevant indicator of bipolar depression. The establishment of the diagnosis of BD is a major challenge and has several consequences: stabilizing the disease, allowing good social reintegration, avoiding relapses and side effects, and, finally, limiting the long-term effects of the disease, particularly on the cognitive level. Treatment strategies usually combine pharmacotherapy (mostly mood stabilizers) and psychosocial care, tailored to each patient. Mood stabilizers aim at decreasing the frequency of major mood episodes. Lithium, some anticonvulsants (such as valproate and carbamazepine), and some antipsychotics (such as aripiprazole, quetiapine, or olanzapine) are the three classes of available mood stabilizers. Psychosocial care includes cognitive rehabilitation strategies, psychoeducation, and interpersonal social and rhythm therapies.

1.3 Schizophrenia

The annual incidence of schizophrenia is 0.2–0.4 per 1000, with a lifetime prevalence of about 0.8% [5], which can slightly vary between countries and cultural groups [6]. These differences are reduced when stricter diagnostic criteria are used for schizophrenia, such as the ones of the DSM-5. Research conducted by the WHO has further confirmed this observation by showing that schizophrenic disorder prevalence is similar across a wide range of cultures and countries, including developed and developing countries [6]. Its sex ratio is around 1:1.

Schizophrenia is characterized by three main types of symptoms, namely, positive symptoms, negative symptoms, and cognitive impairment [7]. Positive symptoms involve a loss of contact with reality; the patient has false beliefs (delusions) and perceptual experiences not shared with others (hallucinations) and may exhibit behavioral oddities. People with schizophrenia can experience different kinds of hallucinations: auditory, visual, olfactory, gustatory, or tactile. About delusions, patients with schizophrenia may have persecutory delusions, control delusions (e.g., belief in telepathy), grandiose delusions (e.g., belief in being a god), and somatic

delusions (e.g., belief that one's body is rotting from the inside) [8]. Negative symptoms are characterized by a deficit state during which basic emotional and behavioral processes are diminished or absent. The most common negative symptoms are blunted affect, anhedonia, avolition, apathy, and alogia (i.e., reduction in the amount or content of speech). Negative symptoms are more frequent and less fluctuating over time than positive symptoms [9]. They are also strongly associated with poor psychosocial functioning [10]. Cognitive impairments in schizophrenia include deficits with attention and concentration, psychomotor speed, learning and memory, and executive function. A decline in cognitive abilities from premorbid functioning is present in most of the patients, with cognitive functioning after the onset of the illness being relatively stable over time [10]. Despite this decline, cognitive functioning in some patients could be in the normal range. As for the negative symptoms, cognitive impairment is strongly associated with poor psychosocial functioning, particularly with regard to social and professional lives.

The etiology of schizophrenia is complex and multifactorial. Genetic and environmental factors seem to play a major role. The risk of developing schizophrenia is higher in patients' relatives than in the general population [11, 12]. Adoption and twin studies have shown that this increased risk is genetic, with the risk being increased by the presence of an affected first-degree relative [12]. There are two main approaches to the treatment of schizophrenia: pharmacological and psychosocial treatments [13]. Antipsychotics constitute the main medication, with major effects on reducing positive symptoms and preventing relapses. First-generation antipsychotics include molecules such as chlorpromazine or haloperidol. Second-generation antipsychotics were developed to decrease the neurological and cognitive side effects. They are the most used molecules nowadays (quetiapine, aripiprazole, risperidone, clozapine, etc.). In contrast, their effects on negative symptoms and cognitive impairment are much more moderate [14]. Psychosocial interventions improve the management of schizophrenia, e.g., through symptom management or relapse prevention. Other specific interventions that can improve the outcome of schizophrenia include family psychoeducation, supported employment, social skills training, psychoeducation, cognitive behavioral therapy, and integrated treatment of comorbid substance abuse [8].

The remainder of this chapter is organized as follows: We first describe the challenges in psychiatry that can potentially be addressed with machine learning. We then provide a non-exhaustive state of the art of machine learning with magnetic resonance imaging in psychiatry. We finally highlight the limitations of current approaches and propose perspectives for the field. Studies reviewed in this chapter are summarized in Table 1.

2 Challenges for Machine Learning in Psychiatry

Diagnosis and treatment are based on clinical diagnostic criteria developed from the subjective human experience, rather than on objective markers of illness. These criteria have been developed based on experts' opinion and are included in the DSM-5 and ICD-10 manuals. This approach has some limitations. Diagnosis can vary across interview methodologies [50], and clinically identical symptoms can be caused by different underlying conditions. Therefore, the common diagnostic criteria, which are based on symptom manifestation alone, are not always reliable in the clinical context [51]. They are indeed often unstable over time and unspecific [52] and provide little guidance to select the appropriate treatment. These misdiagnoses and misclassifications could lead to a poor therapeutic response and suboptimal management of the illness. Based on these observations, it appears necessary to develop objective markers and a better characterization of these illnesses.

In this section, we will discuss how machine learning could be used to improve diagnosis, to help characterize the different mental illnesses, and to improve treatment response and prognostic approach.

2.1 Improving the Diagnosis of Psychiatric Disorders

In the early stages of research on machine learning and psychiatric disorders, researchers wanted to explore whether different diagnoses could be predicted using machine learning algorithms applied to neuroimaging features. They mainly applied machine learning on structural MRI (sMRI) and functional MRI (fMRI) data (during tasks or at rest) [53]. Recent efforts have been made to apply machine learning on diffusion MRI [15], mostly in combination with other modalities [53, 54], and to explore whether adding modalities improves the diagnosis. Classification using machine learning in neuroimaging initially focused on major psychiatric disorders, such as MDD [55], schizophrenia [56], and bipolar disorder [54]. In a second phase, research has broadened the spectrum of psychiatric disorders such as anxiety disorders [23], anorexia [20], substance abuse [57], specific phobia [19], and autism spectrum disorders [58]. Machine learning using EEG has also been investigated for schizophrenia classification [59] as it is an affordable method for functional imaging and since it has a better temporal resolution than fMRI. While lots of machine learning studies in psychiatry focused on neuroimaging data, other fields of research were increasingly interested in using other modalities, such as proteomic, metabolomic [22], and genetic [24] data.

Machine learning also opens perspectives for the identification of relevant features (e.g., the measured variables) for the diagnosis. Using interpretable models such as support vector machines (SVM) or decision trees lets researchers investigate features that are used in

the decision. Deep learning could also be used to find useful features without a priori preprocessing of the images when it is used in combination with interpretation techniques [59]. Another way to identify relevant features for the classification is to compare the prediction performances of different machine learning models with different input features. It then allows us to evaluate if the information present in the different features helps the classification. For example, this was shown in the study of Lin et al. [16], where the authors established that the G72 protein alone yielded almost as much information for the diagnosis of schizophrenia than combined with other G72 single nucleotide polymorphisms. While this approach could be fruitful to build more resilient and interpretable algorithms, we should be careful when interpreting their results. We must keep in mind that statistical algorithms such as the machine learning ones are designed to predict (classes), while inference tests (i.e., univariate statistics) usually rely on association studies, which are more reliable to infer correlation and causal relations [60]. Moreover, when interpreting SVM weights, for example, one must keep in mind that some features are only including noise but are still important when considered in combination with other features [61]. For all these reasons, even though finding important features is necessary to better understand the models, their interpretation to infer pathophysiology or biomarkers must be cautious.

2.2 Refining the Classification of Psychiatric Disorders

Since there is a significant overlap in the clinical symptoms of different psychiatric disorders, many patients suffer from an important delay in the diagnostic establishment, after a potentially harmful diagnosis wavering. For instance, patients with BD wait on average 10 years before receiving an accurate diagnosis [62] and are often misdiagnosed with unipolar depression for years. As for MDD, it is often underdiagnosed even though fast and accurate diagnosis could avoid long-term cognitive impairment in under-treated patients [63]. For all these reasons, making the right diagnosis as early as possible is a major public health challenge.

Machine learning may be a useful tool to discriminate between different diagnoses. Indeed, the interest in machine learning is not only to distinguish a patient with a psychiatric disorder from a healthy subject – which is not the most difficult task for the clinician – but it could be used to help the clinician when the diagnosis becomes more difficult, e.g., to distinguish bipolar depression from unipolar depression [18] or to identify a patient at risk of psychosis [24].

As studies investigate new biomarkers to differentiate between different conditions, our current classification of psychiatric disorders appears to be limited. There are numerous different classification criteria to describe psychopathology, and theoretical frameworks are evolving rapidly [52], which contributes to our

limited understanding of these disorders. The classification of psychiatric disorders is also a complex issue at the biological level, since biological boundaries between conditions are not binary and are blurred by the imprecision of the current genetic and imaging tools (e.g., between BD and schizophrenia [17]). Moreover, the heterogeneity in the clinical presentation of the patients limits the efficiency of a binary classification task. A simple classification algorithm as SVM will only find the largest and shared biomarkers, leading to a suboptimal classification.

The question we might ask is whether changing our perspective and the way we approach psychiatric disorders' heterogeneity will improve our understanding and management of the patients. To consider this heterogeneity, unsupervised machine learning seems to be an appropriate method, as it allows to find new homogeneous subgroup within the population without preconceptions. Current research is using unsupervised machine learning to automatically detect new subgroups (i.e., clusters) of patients based on similar cognitive [25], genetic [64], and/or cerebral [64] profiles. After subgrouping, supervised machine learning can be used to automatically classify the patients into one group or another. For instance, Wu et al. [25] identified two phenotypic groups of patients with BD using a cognitive task battery. Then, they used classifiers to detect white matter tracts' microstructural differences between the two groups. Newly developed algorithms combining supervised learning and clustering show promising results [65], as they can disentangle the heterogeneity of some disorders and improve diagnostic prediction at the same time. The HYDRA model is one of those promising algorithms that has already been used to find some subtypes of Alzheimer disease and to reveal meaningful biomarkers of this disease at the same time [66]. These semi-supervised clustering algorithms [67] are also starting to be used in psychiatry [68] as they could help to reveal biomarkers while discriminating between two different homogeneous classes. Finally, these algorithms are of special interest as they are also handling common source of variation in the groups to be classified (i.e., the age, the sex, or other clinical or biological variables) [69].

Other approaches aim to identify differences between the patients (the cases) and a reference population [70] (the controls). These so-called normative models drop the hypothesis that the patients do not belong to a homogeneous group, which is a step toward a finer analysis. Indeed, recent studies showed important clinical and biological heterogeneity between the patients, especially regarding brain structural abnormalities. Therefore, the hypothesis of an average patient, as it is in classical "case-control" studies, could limit our understanding of the diseases in the long term. Normative modeling could overpass this limitation as it allows to situate a given patient among the "norm" while considering the strong heterogeneity within the patients' population. For

instance, Wolfer et al. [70] showed that deviations from the normative model of gray matter volume were frequent in both SZ and BD but highly heterogeneous. However, these models also induce an asymmetry as they hypothesize that the controls are homogeneous, which is debatable in practice. Nevertheless, it appears that subtyping leads to increased predictive accuracy in identifying individuals with mental illnesses compared with healthy controls, even though results are mixed [71]. This approach could gain attention with the development of new tools such as longitudinal normative brain charts that cover the whole lifespan [72].

2.3 Predicting Evolution and Treatment Response

Predicting the evolution of psychiatric disorders is an important challenge. As previously mentioned, clinicians' choices are guided by recommendations based on broad symptom classifications, such as depression, anxiety, or psychosis criteria, and become personalized over time through an empirical process of trials and errors. Being able to predict the prognosis of the mental illnesses would allow a better organization of care and more adapted psychoeducation consultations, would let clinicians set up strategies to prevent relapses, and would finally greatly improve the quality of life of the patients. Some studies tried to predict psychotic transition using neuroimaging [29] or using EEG [32] and clinical measures [35]. Schmaal et al. [31] used Gaussian process classifiers based on structural and functional MRI (emotional task) to characterize trajectories of depression (chronic, improvement, and rapid remission). They successfully classified the chronic group vs. the rapid remission group with an accuracy of 73%. Regarding other studies on depression, Kessler et al. [73] used self-reported clinical questionnaires of 1057 patients and machine learning algorithm to predict the course of MDD. They predicted the risk of suicide attempt with an AUC of 0.76 and whether the patient would experience a depressive episode lasting more than 2 weeks with an AUC of 0.71. Tran et al. [34] used electronical record's information such as medication, diagnosis, occurrence of interactions with health services, etc. with the aim of stratifying individuals according to their suicide risk. Interestingly, according to their results, their algorithms predicted the suicide risk better than clinicians, with an AUC of 0.73 vs. 0.57 for the prediction of high suicide risk patient vs. the rest of the population. It could also be possible to predict future substance abuse using neuroimaging data [33] and using combinations of demographic, clinical, cognitive, neuroimaging, and genetic data [30]. For schizophrenia, EEG-based machine learning could also be used to determine at-risk patients [59]. Machine learning could also be useful to predict the outcome of a first episode of psychosis [42] and to adapt the treatment. These studies highlight the possibility to stratify and classify individuals to optimize prognostic assessments, thanks to machine

learning. That would help the clinician to propose personalized care, such as primary care facilities for patients at high suicidal risk.

Regarding the treatment outcome, the major challenge is to determine whether machine learning could be used to predict treatment response. This knowledge would be extremely useful, as for now therapeutic choices are made through a trial-error process, which increases the time interval between the apparition of the symptoms and the administration of the adequate treatment. This leads to a serious socioeconomic burden and can have debilitating consequences. In depression, the interest of the machine learning approach was tested on pharmacological decision, for instance, to predict the response to serotonin reuptake inhibitor medications [27]. The authors were able to predict the treatment response using EEG-derived features with an accuracy of 87.9%. In another study, EEG features were also used to predict antipsychotics response in schizophrenia [74]. More recently, studies focused on anatomical and functional MRI. For instance, Whitfield-Gabrieli et al. [28] used resting-state fMRI combined with FA maps as well as initial severity assessment to predict the response to cognitive behavioral therapy in patients with social anxiety. They were able to classify good and poor responders with an accuracy of 81% in a sample of 38 patients. Predicting treatment response is particularly interesting when the treatment is more invasive, such as for the use of electroconvulsive therapy (ECT). Indeed, one team showed (with a sample of 122 depressed patients) that the brain structure can predict the ECT response with an accuracy of 78% [75]. Finally, choosing the right treatment is not just about measuring its effectiveness; it is always about balancing the cost and the acceptable benefit for the patients. In summary, all these features could be integrated in machine learning algorithms and used by the clinicians as tools to improve the accuracy of the therapeutic decisions.

3 MRI and Machine Learning in Psychiatry: State of the Art (Table 1)

To this day, unlike in some medical specialties such as neurology, MRI is rarely used for psychiatric clinical practice. However, it is extensively used in research as it provides a large variety of information about the brain structure and function. Currently, sMRI is the easiest method to implement and the most used in the MRI studies. It is preferentially used to measure the cortex thickness and the cortical surface and to estimate the gray and white matter density and/or volume. Diffusion-weighted imaging (DWI) is less used but provides useful information on the white matter microstructure, thanks to different markers such as fractional anisotropy (the most used), mean diffusivity, and radial diffusivity. fMRI is of particular interest to investigate the neural correlates of cognition and emotion processes and their alteration in patients with

psychiatric conditions. Predictive models are thus useful tools when analyzing MRI data, because they allow to handle high-dimensional inputs and fit more unknown variables than available observations. In neuroimaging, machine learning allows to model sets of effects rather than single effects and thus to build models that describe more than one isolated dimension of cognition.

3.1 Classification Versus Healthy Controls

Classification of patients with psychiatric disorder vs. healthy controls is a widely studied area of research. Even though most studies fail to obtain the 80% of accuracy needed for clinical relevance, they yield promising results and give important methodological insights.

Regarding MDD, using sMRI, machine learning studies [55] found accuracies ranging from 67.6% to 90.3%. These results should be taken with great caution since they are usually obtained from small samples. For example, Mwangi et al. [39] obtained an accuracy of 90.3% using relevance vector machines and a sample of 60 subjects. They also showed that the brain regions identified during the features selection process were consistent with those of previous studies that reported gray matter reductions in patients with MDD, which were mostly located in the frontal lobe, the orbitofrontal and cingulate cortex, the middle frontal gyrus, and the inferior and superior gyri [76]. As for fMRI studies, Gao et al. [55] found an accuracy ranging from 56% to 99%; Ramasubbu et al. [36] found a significant accuracy of 66% for very severe depression using resting-state fMRI in 19 control subjects vs. 45 patients with different intensities of depression; and Fu et al. [21] obtained an accuracy of 86% in a sample of 19 patients with MDD and 19 HC who were processing sad faces during fMRI scanning.

Regarding bipolar disorder (BD), a recent literature review counted 25 studies using machine learning with different MRI modalities to classify BD vs. HC [54]. They found a median accuracy of 66% for BD vs. HC classification. Even though most studies used samples of less than 100 subjects, a study stood out by the number of samples. Using 3040 subjects, sMRI, and a linear SVM, Nunes et al. [43] obtained an accuracy of 65.23% using aggregate subject-level analyses and an accuracy of 58.67% when testing on left out sites. Their results, which highlighted the importance of regions such as the hippocampus, the amygdala, and the inferior frontal gyrus for the classification, were in good accordance with previous MRI studies in BD [76–78]. Regarding fMRI, the review of Claude et al. [54] highlighted that machine learning studies performed with an accuracy range between 37.5% and 83.5%. The minimum accuracy was 37.5% for the classification of bipolar depression vs. HC, during angry face processing using a Gaussian process classifier (GPC) [37]. DWI was not investigated much. In the review of Claude et al. [54], only two DWI studies were referenced. Achalia et al. [15] used DWI and machine learning on 60 subjects and obtained an accuracy of 74% for DWI alone. Even

though DWI gave lower classification scores than sMRI (77.8%) and fMRI (80.3%), combining it with other modalities significantly enhanced the accuracy (87.6%). Mwangi et al. [40] also used DWI in combination with sMRI on 30 pediatric patients with BD and obtained a classification accuracy of 78.12%.

Regarding schizophrenia (SZ), Filippis et al. [56] conducted a systematic review focusing on sMRI and fMRI studies that attempt to classify SZ vs. HC. Notably, the study of Salvador et al. [38] focused on a sample of 128 patients with SZ and 127 HC and aimed to compare the classification score of different neuroimaging features such as voxel-based and wavelet-based (a transformation like Fourier transform) morphometry of gray and white matter, vertex-based cortical thickness and volume defined as regions of interest, as well as volumetric measures. They also compared different methods, such as random forest, regressions with different regularization methods and levels, and SVM. The best results were obtained using the voxel-based and wavelet-based morphometry in combination with a SVM, with respective accuracy of 77.2% and 71%. The authors stress on the fact that no algorithm clearly outperforms the others, but that the selection of features has a real influence on the classification accuracy. Another notable study focused on cortical thickness and surface area measurement to differentiate first-episode psychosis from healthy subjects [42]. This study witnessed that regions contributing to the classification accuracy included the default mode network (DMN), the central executive network, the salience network, and the visual network. They observed a classification accuracy of 85.0% for the surface area and 81.8% for the cortical thickness. Pinaya et al. [79] used a deep belief network, which is a deep neural network that extrapolated and interpreted features, on sMRI data from 83 HC and 143 patients with SZ. The deep belief network highlighted an accuracy of 73.6% vs. 68.1% for a classical SVM. It also detected large differences between classes among specific regions, particularly frontal, temporal, parietal, and insular cortices, the corpus callosum, the putamen, and the cerebellum. Finally, as already mentioned in Subheading 2.1, normative models constructed with MRI data could be a useful tool to handle the inter-subject variability in machine learning models [71].

3.2 Inter-Illness Classification and Clustering

One major challenge of machine learning studies using MRI is to be able to correctly distinguish or classify patients suffering from different disorders. Several studies focused on the classification between BD and SZ. In their review, Claude et al. [54] found that three studies used sMRI in combination with machine learning algorithms to discriminate between BD and SZ with an accuracy ranging between 58% and 66%. Precisely, Schnack et al. [44] showed good classification performance on an independent dataset, with an average classification accuracy of 66%. Mothi et al. [45]

used K-mean clustering after a non-linear PCA to separate patients with BD, SZ, or schizoaffective disorder. They found out that the separation in three clusters was optimal, comprising a cluster including a major proportion of patients with BD, a second with mostly patients with SZ, and a third with a balanced proportion of the three types of illnesses. To build their clusters, they used clinical and cognitive data and validated the robustness of their results with sMRI data. The cluster including more patients with SZ was the one to have a significantly reduced cortical thickness in the frontal lobe. In addition, the BD and the SZ clusters presented significant cortical thickness reductions in occipital and temporal regions.

Several studies attempted to predict the diagnosis of BD in a population of unipolar, bipolar depression, and healthy controls with a median accuracy of 79% and an accuracy ranging from 50% to 90.69% [54]. Burger et al. [37] focused on the classification of unipolar vs. bipolar depression using different regions of interest. They did not find any significant results using the whole brain but found an accuracy of 63.89% for the classification of BD vs. unipolar depression using a GPC based on a happy face processing paradigm and the amygdala activity. Their best accuracy was of 72.2% for the classification of bipolar vs. unipolar depression, using a fear processing paradigm and GPC on the anterior cingulate gyrus. Overall, the best performance was obtained by Grottegerd et al. [18]. In a pilot study, they obtained an accuracy of 90% using fMRI with a happy vs. neutral contrast image and an SVM on 10 BD, 10 HC, and 10 MDD. Using sMRI and DWI with a multiple kernel learning and a sample of 74 MDD and 74 BD, Vai et al. [46] obtained an accuracy of 74.32%, with a positive predictive value of 73.33% (probability that subjects with a positive BD prediction suffer from BD). The accuracy for MDD was 72.97%, indicating the ability to correctly identify people with MDD, with a predictive value of 73.97%. Their models are particularly interesting as they included relevant covariates in their models, such as age, gender, number of previous episodes, and drug load, which can confound and bias the accuracy estimates. Taking into account all these factors helps to increase the performance of the algorithm, as they impact the brain structural measures. It is necessary since these effects were witnessed by the ENIGMA-BD Working Group that used a large cohort of 2447 BD and 4056 HC and found [80] that several commonly prescribed drugs for BD treatment, including lithium, anti-epileptic, and antipsychotic treatments, showed significant associations with cortical thickness and surface area, even after accounting for patients receiving multiple drugs.

3.3 Treatment Response and Illness Prediction

Another perspective is the use of MRI and machine learning algorithms to predict treatment response. This was done by the team of Liu et al. [47] who tested the sensitivity to antidepressants in patients with MDD. Precisely, the study included 17 subjects that

Table 1
Summary of reviewed studies

Study	Modalities	Diseases and goal	Number of subjects	Methods and validation procedure	Results
Achalia et al. [15]	sMRI, dMRI, rs-fMRI, cognitive test	Classification of BD vs. HC	30 HC, 30 BD	SVM and feature selection	Accuracy of 87.60%, sensitivity of 82.3% and specificity of 92.7%
Lin et al. [16]	G72 genotypes, proteins levels	Classification of SZ vs. HC	89 SZ, 60 HC	Naïve Bayes, logistic regression, decision tree	Naïve Bayes model: sensitivity = 0.7969, specificity = 0.9372, (AUC) = 0.9356
Kliemann et al. [17]	Emotion attribution task, localizer task and behavioral task in fMRI	Classification of Autism spectrum disorder (ASD) vs. HC	Study 1: ASD: 16, HC: 21 Study 2: ASD: 22, HC: 30	ROI-based MVPA (Multivariate pattern analysis) with a SVM	Only statistics on the MVPA results
Grotgerd et al. [18]	Facial emotion processing, fMRI	Classification of BD vs. HC vs. Unipolar depression (UD)	BD: 10, HC: 10, UD: 10	SVM, GPC	Best results: happy versus neutral with the SVM for UD vs. BD: 90% accuracy
Visser et al. [19]	Generalization task, fMRI	Identifying area that influence the fear of spiders in spider-phobia stricken patients	High spider fear patients (HSF): 18, Low spider fear patient (LSF): 20	SVM-based MVPA	Several above chance results of correct classification of the stimuli depending on the ROIs
Lavagetto et al. [20]	sMRI	Classification of Anorexia Nervosa (AN) patient versus HC	AN: 15, HC: 15	LASSO linear regression	83.3% accuracy (sensitivity 86.7%, specificity 80.0%)
Fu et al. [21]	Sad face processing, fMRI	Classification of MDD versus HC	MDD: 19, HC: 19	SVM	Accuracy of 86% (84% of sensitivity and 89% of specificity)

(continued)

Table 1
(continued)

Study	Modalities	Diseases and goal	Number of subjects	Methods and validation procedure	Results
Setoyama et al. [22]	Plasma metabolites concentration	Prediction of suicide ideation (SI) in MDD and BD patients	Dataset 1: 51 drug-free MDD patients Dataset 2: 23 medicated MDD patients Dataset 3: 27 MDD, 14 BD	Logistic regression, SVM and random forest. Results reported for each fold of a 10-fold cross validation	Logistic regression vs. no SI: 57.14% (mean of the 10-fold) SVM accuracy of SI vs. no SI: 52.08% (mean of the 10-fold) Random forest: overfit
Lueken et al. [23]	Fear conditioning task, fMRI	Panic disorder, MDD	Panic disorder: 33, Panic disorder with depression: 26	Adaboost	Comorbidity status was correctly predicted in 79% of patients (sensitivity: 73%; specificity: 85%) based on brain activation during fear conditioning (corrected for potential confounders: accuracy: 73%; sensitivity: 77%; specificity: 70%)
Pettersson-Yeo et al. [24]	Genotype, sMRI, dMRI, cognitive data, fMRI	Classification of ultra high risk (UHR) patients vs. first episode psychosis patient (FEP) vs. HC	HC: 23, UHR: 19, FEP: 19	SVM	FEP vs. HC (genotype, 67.86%; dMRI, 65.79%; fMRI, 65.79% and 68.42%; cognitive data, 73.69%), UHR versus HC (sMRI, 68.42%; dMRI, 65.79%), and FEP

				versus UHR (sMRI, 76.67% fMRI, 73.33%; cognitive data, 66.67%).
Wu et al. [25]	Neurocognitive data, dMRI (FA, MD)	Unsupervised clustering of BD patients confirmed by supervised classification	K-means, ElasticNet	Identification of two phenotypes by K-means
Drysdale et al. [26]	rsfMRI	Unsupervised clustering of depressive disorder	Dataset 1: 333 DD, 378 HC Dataset 2: 352 HC, 109 DD, 16 BD II	K-means, SVM, Logistic regression, LDA with cross validation
Khodayari- Rostamabad et al. [27]	EEG	Prediction of the response to SSRI antidepressant therapy in treatment- resistant MDD	Mixture of factor analysis (MFA)	Specificity: 80.9%, sensitivity: 94.9%, accuracy: 87.9%.
Whitfield-Gabriel et al. [28]	Rs-fMRI, dMRI	Prediction of cognitive behavioral therapy results on social anxiety disorder inpatients	38 social anxiety disorder patients	Logistic regression and MVPA, leave-one out cross validation
Koutsouleris et al. [29]	sMRI	Classification of SZ vs. MDD	Database 1: 104 MDD, 158 SZ, 437 HC Database 2: 35 BD, 23 FEP, 89 At risk patients	Classification of MD vs. SZ: 76% of balanced accuracy.

(continued)

Table 1
(continued)

Study	Modalities	Diseases and goal	Number of subjects	Methods and validation procedure	Results
The IMAGEN Consortium et al. [30]	Stop-signal task, fMRI, personality measure, cognitive test	Prediction of future alcohol-misuse	692 teenagers	ElasticNet logistic regression	94% of correct classification of binge-drinkers and 34% of non binge-drinkers at the optimum of the Receiver Operating Characteristic curve.
Schmaal et al. [31]	Face task and London tower task, fMRI, clinical data, sMRI	Prediction of MDD prognosis	156 MDD, 145 HC	GPC	fMRI: 73% of correct classification for chronic patients vs. more favorable trajectories patients. 69% of accuracy for chronic vs. remitted patients with clinical data.
Ramyead et al. [32]	EEG	Prediction of psychosis transition	53 at risk of psychosis inpatients (18 made a transition to psychosis)	LASSO regression	Prediction of psychotic transition: AUC of 0.77
Bertocci et al. [33]	reward-task, fMRI	Prediction of substance abuse in emotionally dysregulated teenagers	30 emotionally dysregulated female teenagers	LASSO regression	Classification of substance abuse vs. no substance abuse: 83.6%

Tran et al. [34]	Electronic medical record (EMR)	Risk stratification of suicidal behaviour	7399 subjects with risk assessment	L1 continuation-ratio model for ordinal outcomes	Classification of high-risk patients versus the rest using EMR data: AUC of 0.79
Mechelli et al. [35]	Clinical data	Psychotic disorders	416 subjects at risk for psychosis	SVM with leave-one-out cross validation	64.6% of accuracy for the prognostic prediction of transition to psychosis
Ramasubbu et al. [36]	Resting-state fMRI, Emotional-faces task fMRI	Classification of different MDD subgroups vs. HC	45 MDD, 19 HC	ICA for feature selection SVM for classification	Very severe group vs. HC with rs-fMRI: 66% accuracy Rest of the classifications at chance levels
Bürger et al. [37]	Emotional faces task fMRI.	Classification of BD vs. HC vs. UD Features from the anterior cingulate gyrus or the amygdala	36 HC, 36 BD, 36 UD	SVM or GPC for the MVPA	Significant results BD vs. UC with a SVM and fearful faces: 69.44%, with GPC: 72%. UC vs. HC with GPC and happy faces: 66.67%.
Salvador et al. [38]	sMRI features: vertex-based cortical thickness, grey and white matter voxels/wavelet-based morphometry, regional volumes and interactions.	Classification of HC vs. SZ vs. BD	127 HC, 128 BD, 128 SZ	Logistic regression, Lasso, Ridge and ElasticNet, SVC, RDA, GPC, Random forest	Best results with VBM. SZ vs. HC: 75%, BD vs. HC: 63% and SZ vs. BD: 62%.
Mwangi et al. [39]	sMRI features	Classification of MDD	30 MDD, 32 HC	SVM and Relevance Vector Machine, train/test split	Classification of 90.3% for MDD vs. HC, correlation of Relevance Vector Machine prediction and illness severity

(continued)

Table 1
(continued)

Study	Modalities	Diseases and goal	Number of subjects	Methods and validation procedure	Results
Mwangi et al. [40]	dMRI features	Prediction of diagnosis in pediatric BD	16 pediatric BD patients, 16 matched HC	SVM with leave-one-out cross-validation and grid search	The SVM algorithm trained with a linear kernel function and FA values performed best with accuracy = 78.12%, sensitivity = 68.75%, specificity = 87.5%
Costafreda et al. [41]	sMRI	Classification of MDD vs. HC and prognosis prediction of CBT and fluoxetine treatment	37 MDD, 37 HC	SVM and leave-one-out cross-validation	The whole brain structural neuroanatomy predicted 88.9% of the clinical response. Accuracy of the structural neuroanatomy as a diagnostic marker was 67.6%.
Xiao et al. [42]	sMRI features	Classification of first episode psychosis patients vs. HC	163 drug naïve first episode psychosis patients, 163 HC	SVM and 10-fold cross-validation	Surface Area. Accuracy: 85.0% (specificity = 87.0%, sensitivity = 83.0%) Cortical thickness. Accuracy: 81.8% (specificity = 85.0%, sensitivity = 76.9%)

Abraham et al. for the ENIGMA Bipolar Disorders Working Group et al. [43]	sMRI features	Classification of BD vs. HC HC	853 BD, 2167 HC	SVM and 30-fold cross-validation or leave-one-site-out	Aggregate subject analysis: accuracy = 65.23%
Schnack et al. [44]	sMRI features (GMD)	Classification of BD vs. SZ vs. HC	112 SZ, 113 BD and 109 HC	3 SVM models (HC vs. BD, HC vs. SZ and BD vs. SZ) tested on a validation sample	Leave-one-site-out cross-validation: accuracy = 58.67%
Mothi et al. [45]	Clinical measures, eye movement tracks, cognitive assessment, sMRI	Unsupervised clustering of BD, SZ and schizoaffective patients	251 SZ, 164 SZA, 195 BD	K-means with 10 first components of a PCA ran on clinical measures, eye movement, cognitive assessment	On the validation sample: ROC AUC = 66% for BD vs. SZ, 55.5% for BD vs. HC, 79.2% for SZ vs. HC
Vai et al. [46]	sMRI and dMRI features	Classification of BD vs. MDD (FA)	74 BD, 74 MDD	MVPA with 10-fold cross-validation	3 clusters with one with a majority of SZ, one with a majority of BD, and one distributed across the 3 disorders
Liu et al. [47]	sMRI (VBM of GM or WM)	Classification of treatment resistant depression vs. treatment sensitive depression	17 TRD, 17 TSD, 17 HC	MVPA with leave-one-out cross-validation	Balanced accuracy of 73.65% with a sensitivity for BD of 74.32% and specificity for MDD of 72.97%
Hajek et al. [48]	sMRI (GM and WM separately)	Classification of unaffected high-risk offspring (UHRO) versus controls (C)	45 UHRO, 45 C	SVM, GPC, leave-2-out cross-validation	GM results: TRD vs. TSD: 82.9%, TRD vs. HC 85.7%, TSD vs. HC 82.4%. WM results: TRD vs. TSD 82.9, TRD vs. HC 85.7%, TSD vs. HC 91.2%

(continued)

Table 1
(continued)

Study	Modalities	Diseases and goal	Number of subjects	Methods and validation procedure	Results
Fleck et al. [49]	fMRI (continuous performance task with distractions), proton magnetic resonance spectroscopy (PMRS)	Predicting treatment response	20 BD that received lithium treatment	Cascading genetic fuzzy tree	80% accuracy

Acronyms. *AN* anorexia nervosa, *ASD* autism spectrum disorder, *BD* bipolar disorder, *C* controls, *DD* depressive disorder, *dMRI* diffusion MRI, *FA* fractional anisotropy, *FEP* first episode psychosis, *fMRI* functional MRI, *GM* gray matter, *GMD* gray matter density, *GPC* Gaussian process classifier, *HC* healthy controls, *ICA* independent component analysis, *MD* mean diffusivity, *MDD* major depressive disorder, *MR* magnetic resonance imaging, *MVPA* multivariate pattern analysis, *PMRS* proton magnetic resonance spectroscopy, *rfdMRI* resting-state functional MRI, *sMRI* structural MRI, *SI* suicide ideation, *SVM* support vector machine, *SZA* schizoaffective, *TRD* treatment-resistant depression, *TSD* treatment-sensitive depression, *UD* unipolar depression, *VBM* voxel-based morphometry, *WM* white matter

were treatment resistant, 17 that were treatment sensitive, and 17 controls. The accuracy of the MVPA models that correctly distinguished resistant and sensitive patients from HC ranged from 85.7% to 91.2% depending on the features used. The authors highlighted differences in structural alterations between responders and non-responders suggesting that structural differences may be related to different responses to antidepressants. Furthermore, they found that the structural abnormalities were larger between responders and HC than between non-responders and HC. These results are somewhat counterintuitive as one would expect resistant patients to show more structural differences from HC than responders. However, this lack of specificity is probably related to a high degree of clinical heterogeneity and the small sample size that does not allow sufficient precision to distinguish more specific abnormalities.

Hajek et al. [48] used machine learning applied to white matter sMRI to distinguish 45 unaffected participants at high genetic risk of BD from 45 low-risk healthy controls with an accuracy of 68.9%. Similarly, Lin et al. [81] successfully classified HR individuals for BD with vs. without (sub)syndromic risk with an accuracy of 83.21% based on the gray matter volume. Finally, a pilot study was conducted using a novel machine learning system based on a “multi-cascade fuzzy genetic tree” with sMRI capable of accurately classifying subjects with BD in a first manic episode into groups that responded or did not respond to lithium treatment [49].

4 Limitations and Perspectives

As illustrated in this chapter, numerous studies have been conducted to classify psychiatric disorders and refine the definition of psychiatric subgroups using machine learning. However, methods and results are heterogeneous. In fact, many authors point to a major limitation of most studies, that is, the limited number of samples [53–55]. Claude et al. [54] also pointed out a negative correlation between the accuracy and the number of subjects, leading to think that the results obtained from small samples are artificially high. Another effect of this limited number of samples resides on the fact that models need to be trained on a population that is representative of the population on which we will use them. Indeed, models trained on a young population will be biased when used on an older one, and similar bias could be raised when using a model trained with a population from a specific country on subjects from another country.

As it is difficult to recruit enough patients to obtain a sufficient statistical power, this limitation may persist in the long term, unless

collective efforts for data sharing are undertaken. This issue deepens when looking at more specific subsets of patients. The field therefore needs more and larger datasets to work on. These datasets start to be collected, with, e.g., the UK BioBank dataset (~40,000 subjects). Even though they are not focused on psychiatric disorders, they are interesting because they are multimodal datasets, with genetic, clinical, and MRI data, and some participants will develop psychiatric syndromes throughout the follow-up. Recent efforts have been specifically made for psychiatric disorders, e.g., by the ENIGMA Consortium, a multisite and multimodal project including several working groups focused on different diseases, such as bipolar disorder, schizophrenia, autism, ADHD, etc.

Larger datasets are often multisite ones, and they bring their own challenges. Since the MRI devices that are used for different studies have different magnetic field strengths, different vendors, coils, etc., there are large site effects that need to be considered. These site effects are particularly important for DWI and fMRI, but they even appear for sMRI [82], the most robust method of imaging. A second source of site effects lies in the preprocessing of the data, which may vary between different sites and protocols. The preprocessing steps are of major importance and need to be homogenized since different softwares can lead to different results [83]. The remaining “site effects” can be partially corrected, thanks to different methods. Statistics-based methods include adjusted residualizations or ComBat [84, 85], a method originally proposed to remove batch effects in genomics [86] and then adapted for DWI and then for sMRI [87]. Other methods are more specific to MRI, such as RAVEL [88], which aims at capturing the sites’ variability using the signal from the CSF, with mixed results for now. Since the extent of the efficiency of these corrections is still under discussion [89], we must consider the site effect in our models and use validation methods such as leave-one-site-out validation to evaluate the reproducibility of our approaches.

The site effect highlights a deeper and more fundamental limitation of our studies, the signal-to-noise ratio. That issue, which is faced by all imaging studies, is particularly present in neuroimaging for psychiatric diseases as the changes that we are looking for are subtle and probably not the main causes of variation in our datasets (e.g., one important cause of variance is the age, which produces consequent variations in the gray and white matter density [72]). We therefore need to be vigilant and make specific efforts when interpreting the results of machine learning algorithms as they can learn some information that are irrelevant for psychiatric disorders. Nevertheless, it is possible to improve this signal-to-noise ratio. One way to do so is to improve the signal; the second is to diminish the noise. Larger datasets improve the statistical power of the algorithms but may induce noise (such as the multisite noise). In addition to the fact that methodological modifications can change

and improve the performance of machine learning, technological improvements seem to bring better performance as shown by the team of Iwabuchi et al. [90], who showed that 7 T MRI compared to 3 T MRI gave higher classification accuracy when distinguishing patients with schizophrenia vs. controls (77% versus 66%). Moreover, the use of multimodal datasets has shown promising results in increasing the signal-to-noise ratio in current studies [91]. While trying to determine to what extent machine learning using MRI can still improve its results, Schulz et al. [91] highlighted two interesting perspectives: first, that there is still room for improvement of the classification accuracy by getting larger datasets and second, that multimodal MRI and more specifically fMRI could improve the classification.

Other ways to collect data could also be thought about, with, for example, the use of tools such as smartphones. Data can be provided through active monitoring (self-reporting), passive monitoring of various activities, mobility, or statistics on phone calls [92]. Promising results show that voice data from daily phone calls could be a valid marker of mood states and hold promise for monitoring BD [93]. Taken together, the development of our knowledge of machine learning and the growing data resources could provide new tools for the management of psychiatric disorders soon. However, their development can only be done by considering the challenges they raise, such as personal data protection, but also by considering all the ethical issues that these new tools will raise.

Finally, machine learning in psychiatry is a promising field of research, with still a lot to do to characterize the different biomarkers and psychiatric disorders properly and accurately. The use of MRI and other clinical and biological features could in a near future bring new tools for diagnosis, risk assessment, and treatment selection that could be used by the clinician. However, due to the actual social stigma around psychiatric disorders and the apparent arbitrary character of classification algorithms, their use would need an important ethical discussion beforehand, notably when people would like to use them to identify at-risk healthy subjects or when using them to determine the treatment of already symptomatic patients.

Acknowledgments

We are grateful to the reviewer, Anton Iftimovici, for his very helpful comments and suggestions.

References

1. Kupfer DJ, Frank E, Phillips ML (2012) Major depressive disorder: new clinical, neurobiological, and treatment perspectives. *Lancet* 379(9820):1045–1055. [https://doi.org/10.1016/S0140-6736\(11\)60602-8](https://doi.org/10.1016/S0140-6736(11)60602-8)
2. Kessler RC, Angermeyer M, Anthony JC et al (2007) Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's world mental health survey initiative. *World Psychiatry* 6(3): 168–176
3. Diagnostic and statistical manual of mental disorders: DSM-5™, 5th ed. American Psychiatric Publishing, Inc. 2013. p. xliv, 947. <https://doi.org/10.1176/appi.books.9780890425596>
4. Grande I, Berk M, Birmaher B, Vieta E (2016) Bipolar disorder. *Lancet* 387(10027): 1561–1572. [https://doi.org/10.1016/S0140-6736\(15\)00241-X](https://doi.org/10.1016/S0140-6736(15)00241-X)
5. Jablensky A (1997) The 100-year epidemiology of schizophrenia. *Schizophr Res* 28(2–3): 111–125. [https://doi.org/10.1016/S0920-9964\(97\)85354-6](https://doi.org/10.1016/S0920-9964(97)85354-6)
6. Institute of Medicine (US) Committee on Nervous System Disorders in Developing Countries (2001) Neurological, psychiatric, and developmental disorders: meeting the challenge in the developing world. National Academies Press (US). Accessed September 10, 2021. <http://www.ncbi.nlm.nih.gov/books/NBK223475/>
7. Mueser KT, McGurk SR (2004) Schizophrenia. *Lancet* 363(9426):2063–2072. [https://doi.org/10.1016/S0140-6736\(04\)16458-1](https://doi.org/10.1016/S0140-6736(04)16458-1)
8. Fenton WS (1991) Natural history of schizophrenia subtypes: II. Positive and negative symptoms and long-term course. *Arch Gen Psychiatry* 48(11):978. <https://doi.org/10.1001/archpsyc.1991.01810350018003>
9. Sayers SL, Curran PJ, Mueser KT (1996) Factor structure and construct validity of the scale for the assessment of negative symptoms. *Psychol Assess* 8(3):269–280. <https://doi.org/10.1037/1040-3590.8.3.269>
10. Green MF, Kern RS, Braff DL, Mintz J (2000) Neurocognitive deficits and functional outcome in schizophrenia: are we measuring the “right stuff”? *Schizophr Bull* 26(1):119–136. <https://doi.org/10.1093/oxfordjournals.schbul.a033430>
11. McGuffin P, Owen Michael J, Farmer Anne E (1995) Genetic basis of schizophrenia. *Lancet* 346(8976):678–682. [https://doi.org/10.1016/S0140-6736\(95\)92285-7](https://doi.org/10.1016/S0140-6736(95)92285-7)
12. Cardno AG, Marshall EJ, Coid B et al (1999) Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch Gen Psychiatry* 56(2):162. <https://doi.org/10.1001/archpsyc.56.2.162>
13. Davis JM, Chen N, Glick ID (2003) A meta-analysis of the efficacy of second-generation antipsychotics. *Arch Gen Psychiatry* 60(6): 553. <https://doi.org/10.1001/archpsyc.60.6.553>
14. Insel TR, Cuthbert BN (2015) Brain disorders? Precisely. *Science* 348(6234):499–500. <https://doi.org/10.1126/science.aab2358>
15. Achalia R, Sinha A, Jacob A et al (2020) A proof of concept machine learning analysis using multimodal neuroimaging and neurocognitive measures as predictive biomarker in bipolar disorder. *Asian J Psychiatr* 50:101984. <https://doi.org/10.1016/j.ajp.2020.101984>
16. Lin E, Lin CH, Lai YL, Huang CH, Huang YJ, Lane HY (2018) Combination of G72 genetic variation and G72 protein level to detect schizophrenia: machine learning approaches. *Front Psych* 9:566. <https://doi.org/10.3389/fpsyg.2018.00566>
17. Yamada Y, Matsumoto M, Iijima K, Sumiyoshi T (2020) Specificity and continuity of schizophrenia and bipolar disorder: relation to biomarkers. *Curr Pharm Des* 26(2):191–200. <https://doi.org/10.2174/13816282566191216153508>
18. Grotegerd D, Suslow T, Bauer J et al (2013) Discriminating unipolar and bipolar depression by means of fMRI and pattern classification: a pilot study. *Eur Arch Psychiatry Clin Neurosci* 263(2):119–131. <https://doi.org/10.1007/s00406-012-0329-4>
19. Visser RM, Haver P, Zwitser RJ, Scholte HS, Kindt M (2016) First steps in using multi-voxel pattern analysis to disentangle neural processes underlying generalization of spider fear. *Front Hum Neurosci* 10. <https://doi.org/10.3389/fnhum.2016.00222>
20. Lavagnino L, Amianto F, Mwangi B et al (2015) Identifying neuroanatomical signatures of anorexia nervosa: a multivariate machine learning approach. *Psychol Med* 45(13): 2805–2812. <https://doi.org/10.1017/S0033291715000768>
21. Fu CHY, Mourao-Miranda J, Costafreda SG et al (2008) Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. *Biol Psychiatry* 63(7):656–662. <https://doi.org/10.1016/j.biopsych.2007.08.020>

22. Setoyama D, Kato TA, Hashimoto R et al (2016) Plasma metabolites predict severity of depression and suicidal ideation in psychiatric patients – a multicenter pilot analysis. Hashimoto K, ed. PLoS One 11(12): e0165267. <https://doi.org/10.1371/journal.pone.0165267>
23. Lueken U, Straube B, Yang Y et al (2015) Separating depressive comorbidity from panic disorder: a combined functional magnetic resonance imaging and machine learning approach. J Affect Disord 184:182–192. <https://doi.org/10.1016/j.jad.2015.05.052>
24. Pettersson-Yeo W, Benetti S, Marquand AF et al (2013) Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. Psychol Med 43(12): 2547–2562. <https://doi.org/10.1017/S003329171300024X>
25. Wu MJ, Mwangi B, Bauer IE et al (2017) Identification and individualized prediction of clinical phenotypes in bipolar disorders using neurocognitive data, neuroimaging scans and machine learning. NeuroImage 145:254–264. <https://doi.org/10.1016/j.neuroimage.2016.02.016>
26. Drysdale AT, Gosenick L, Downar J et al (2017) Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 23(1):28–38. <https://doi.org/10.1038/nm.4246>
27. Khodayari-Rostamabad A, Reilly JP, Hasey GM, de Bruin H, MacCrimmon DJ (2013) A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. Clin Neurophysiol 124(10):1975–1985. <https://doi.org/10.1016/j.clinph.2013.04.010>
28. Whitfield-Gabrieli S, Ghosh SS, Nieto-Castanon A et al (2016) Brain connectomics predict response to treatment in social anxiety disorder. Mol Psychiatry 21(5):680–685. <https://doi.org/10.1038/mp.2015.109>
29. Koutsouleris N, Meisenzahl EM, Borgwardt S et al (2015) Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. Brain 138(7): 2059–2073. <https://doi.org/10.1093/brain/awv111>
30. The IMAGEN Consortium, Whelan R, Watts R et al (2014) Neuropsychosocial profiles of current and future adolescent alcohol misusers. Nature 512(7513):185–189. <https://doi.org/10.1038/nature13402>
31. Schmaal L, Marquand AF, Rhebergen D et al (2015) Predicting the naturalistic course of major depressive disorder using clinical and multimodal neuroimaging information: a multivariate pattern recognition study. Biol Psychiatry 78(4):278–286. <https://doi.org/10.1016/j.biopsych.2014.11.018>
32. Ramyead A, Studerus E, Komter M et al (2016) Prediction of psychosis using neural oscillations and machine learning in neuroleptic-naïve at-risk patients. World J Biol Psychiatry 17(4):285–295. <https://doi.org/10.3109/15622975.2015.1083614>
33. Bertocci MA, Bebko G, Versace A et al (2017) Reward-related neural activity and structure predict future substance use in dysregulated youth. Psychol Med 47(8):1357–1369. <https://doi.org/10.1017/S0033291716003147>
34. Tran T, Luo W, Phung D et al (2014) Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. BMC Psychiatry 14(1):76. <https://doi.org/10.1186/1471-244X-14-76>
35. Mechelli A, Lin A, Wood S et al (2017) Using clinical information to make individualized prognostic predictions in people at ultra high risk for psychosis. Schizophr Res 184:32–38. <https://doi.org/10.1016/j.schres.2016.11.047>
36. Ramasubbu R, Brown MRG, Cortese F et al (2016) Accuracy of automated classification of major depressive disorder as a function of symptom severity. NeuroImage Clin 12:320–331. <https://doi.org/10.1016/j.nicl.2016.07.012>
37. Bürger C, Redlich R, Grotegerd D et al (2017) Differential abnormal pattern of anterior cingulate gyrus activation in unipolar and bipolar depression: an fMRI and pattern classification approach. Neuropsychopharmacology 42(7): 1399–1408. <https://doi.org/10.1038/npp.2017.36>
38. Salvador R, Radua J, Canales-Rodríguez EJ et al (2017) Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. Hu D, ed. PLoS One 12(4):e0175683. <https://doi.org/10.1371/journal.pone.0175683>
39. Mwangi B, Ebmeier KP, Matthews K, Douglas SJ (2012) Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. Brain 135(5):1508–1521. <https://doi.org/10.1093/brain/aws084>
40. Mwangi B, Wu MJ, Bauer IE et al (2015) Predictive classification of pediatric bipolar disorder using atlas-based diffusion weighted imaging and support vector machines. Psychiatry Res Neuroimaging 234(2):265–271.

- <https://doi.org/10.1016/j.pscychresns.2015.10.002>
41. Costafreda SG, Chu C, Ashburner J, Fu CHY (2009) Prognostic and diagnostic potential of the structural neuroanatomy of depression. *PLoS One* 4(7):e6353. <https://doi.org/10.1371/journal.pone.0006353>
 42. Xiao Y, Yan Z, Zhao Y et al (2019) Support vector machine-based classification of first episode drug-naïve schizophrenia patients and healthy controls using structural MRI. *Schizophr Res* 214:11–17. <https://doi.org/10.1016/j.schres.2017.11.037>
 43. ENIGMA Bipolar Disorders Working Group, Nunes A, Schnack HG et al (2020) Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA bipolar disorders working group. *Mol Psychiatry* 25(9):2130–2143. <https://doi.org/10.1038/s41380-018-0228-9>
 44. Schnack HG, Nieuwenhuis M, van Haren NEM et al (2014) Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. *NeuroImage* 84:299–306. <https://doi.org/10.1016/j.neuroimage.2013.08.053>
 45. Mothi SS, Sudarshan M, Tandon N et al (2019) Machine learning improved classification of psychoses using clinical and biological stratification: update from the bipolar-schizophrenia network for intermediate phenotypes (B-SNIP). *Schizophr Res* 214:60–69. <https://doi.org/10.1016/j.schres.2018.04.037>
 46. Vai B, Parenti L, Bollettini I et al (2020) Predicting differential diagnosis between bipolar and unipolar depression with multiple kernel learning on multimodal structural neuroimaging. *Eur Neuropsychopharmacol* 34:28–38. <https://doi.org/10.1016/j.euroneuro.2020.03.008>
 47. Liu F, Guo W, Yu D et al (2012) Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. Fan Y, ed. *PLoS One* 7(7):e40968. <https://doi.org/10.1371/journal.pone.0040968>
 48. Hajek T, Cooke C, Kopecek M, Novak T, Hoschl C, Alda M (2015) Using structural MRI to identify individuals at genetic risk for bipolar disorders: a 2-cohort, machine learning study. *J Psychiatry Neurosci* 40(5):316–324. <https://doi.org/10.1503/jpn.140142>
 49. Fleck DE, Ernest N, Adler CM et al (2017) Prediction of lithium response in first-episode mania using the LITHium intelligent agent (LITHIA): pilot data and proof-of-concept. *Bipolar Disord* 19(4):259–272. <https://doi.org/10.1111/bdi.12507>
 50. Miller PR, Dasher R, Collins R, Griffiths P, Brown F (2001) Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews. *Psychiatry Res* 105:255
 51. ECNP consensus meeting. Bipolar depression. Nice, March 2007|<https://doi.org/10.1016/j.euroneuro.2008.03.003>. Published September 8, 2021
 52. Alda M (2021) The moving target of psychiatric diagnosis. *J Psychiatry Neurosci* 46(3):E415–E417. <https://doi.org/10.1503/jpn.210098>
 53. Arbabshirani MR, Plis S, Sui J, Calhoun VD (2017) Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *NeuroImage* 145:137–165. <https://doi.org/10.1016/j.neuroimage.2016.02.079>
 54. Claude L, Houenou J, Duchesnay E, Favre P (2020) Will machine learning applied to neuroimaging in bipolar disorder help the clinician? A critical review and methodological suggestions. *Bipolar Disord* 22(4):334–355. <https://doi.org/10.1111/bdi.12895>
 55. Gao S, Calhoun VD, Sui J (2018) Machine learning in major depression: from classification to treatment outcome prediction. *CNS Neurosci Ther* 24(11):1037–1052. <https://doi.org/10.1111/cns.13048>
 56. de Filippis R, Carbone EA, Gaetano R et al (2019) Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: a systematic review. *Neuropsychiatr Dis Treat* 15:1605–1627. <https://doi.org/10.2147/NDT.S202418>
 57. Kruschwitz JD, Ludwig VU, Waller L et al (2018) Regulating craving by anticipating positive and negative outcomes: a multivariate pattern analysis and network connectivity approach. *Front Behav Neurosci* 12:297. <https://doi.org/10.3389/fnbeh.2018.00297>
 58. Kliemann D, Richardson H, Anzellotti S et al (2018) Cortical responses to dynamic emotional facial expressions generalize across stimuli, and are sensitive to task-relevance, in adults with and without autism. *Cortex* 103:24–43. <https://doi.org/10.1016/j.cortex.2018.02.006>
 59. Barros C, Silva CA, Pinheiro AP (2021) Advanced EEG-based learning approaches to predict schizophrenia: promises and pitfalls. *Artif Intell Med* 114:102039. <https://doi.org/10.1016/j.artmed.2021.102039>

60. Bzdok D, Ioannidis JPA (2019) Exploration, inference, and prediction in neuroscience and biomedicine. *Trends Neurosci* 42(4):251–262. <https://doi.org/10.1016/j.tins.2019.02.001>
61. Haufe S, Meinecke F, Görzen K et al (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* 87:96–110. <https://doi.org/10.1016/j.neuroimage.2013.10.067>
62. Bowden CL (2001) Strategies to reduce misdiagnosis of bipolar depression. *Psychiatr Serv* 52(1):51–55. <https://doi.org/10.1176/appi.ps.52.1.51>
63. Galimberti C, Bosi MF, Volontè M, Giordano F, Dell'Osso B, Viganò CA (2020) Duration of untreated illness and depression severity are associated with cognitive impairment in mood disorders. *Int J Psychiatry Clin Pract* 24(3):227–235. <https://doi.org/10.1080/13651501.2020.1757116>
64. Arnedo J, Svrakic DM, del Val C et al (2015) Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome-wide association studies. *Am J Psychiatry* 172(2):139–153. <https://doi.org/10.1176/appi.ajp.2014.14040435>
65. Schnack HG (2019) Improving individual predictions: machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophr Res* 214:34–42. <https://doi.org/10.1016/j.schres.2017.10.023>
66. Varol E, Sotiras A, Davatzikos C (2017) HYDRA: revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. *NeuroImage* 145:346–364. <https://doi.org/10.1016/j.neuroimage.2016.02.041>
67. Schulz MA, Chapman-Rounds M, Verma M, Bzdok D, Georgatzis K (2020) Inferring disease subtypes from clusters in explanation space. *Sci Rep* 10(1):12900. <https://doi.org/10.1038/s41598-020-68858-7>
68. Yang T, Frangou S, Lam RW et al (2021) Probing the clinical and brain structural boundaries of bipolar and major depressive disorder. *Transl Psychiatry* 11:48. <https://doi.org/10.1038/s41398-020-01169-7>
69. Louiset R, Gori P, Dufumier B, Houenou J, Grigis A, Duchesnay E. UCSL: a machine learning expectation-maximization framework for unsupervised clustering driven by supervised learning. *ArXiv210701988 Cs Stat*. Published online July 5, 2021. Accessed March 1, 2022. <http://arxiv.org/abs/2107.01988>
70. Wolfers T, Doan NT, Kaufmann T et al (2018) Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. *JAMA Psychiatry* 75(11):1146. <https://doi.org/10.1001/jamapsychiatry.2018.2467>
71. Marquand AF, Rezek I, Buitelaar J, Beckmann CF (2016) Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biol Psychiatry* 80(7):552–561. <https://doi.org/10.1016/j.biopsych.2015.12.023>
72. Bethlehem RAI, Seidlitz J, White SR et al (2022) Brain charts for the human lifespan. *Nature* 604(7906):525–533. <https://doi.org/10.1038/s41586-022-04554-y>
73. Kessler RC, van Loo HM, Wardenaar KJ et al (2016) Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Mol Psychiatry* 21(10):1366–1371. <https://doi.org/10.1038/mp.2015.198>
74. Suvisaari J, Mantere O, Keinänen J et al (2018) Is it possible to predict the future in first-episode psychosis? *Front Psych* 9:580. <https://doi.org/10.3389/fpsyg.2018.00580>
75. Sun H, Jiang R, Qi S et al (2019) Preliminary prediction of individual response to electroconvulsive therapy using whole-brain functional magnetic resonance imaging data. *Neuroimage Clin* 26:102080. <https://doi.org/10.1016/j.nicl.2019.102080>
76. Hajek T, Cullis J, Novak T et al (2013) Brain structural signature of familial predisposition for bipolar disorder: replicable evidence for involvement of the right inferior frontal gyrus. *Biol Psychiatry* 73(2):144–152. <https://doi.org/10.1016/j.biopsych.2012.06.015>
77. Hajek T, Kopecek M, Kozeny J, Gunde E, Alda M, Höschl C (2009) Amygdala volumes in mood disorders — meta-analysis of magnetic resonance volumetry studies. *J Affect Disord* 115(3):395–410. <https://doi.org/10.1016/j.jad.2008.10.007>
78. Ganzola R, Duchesne S (2017) Voxel-based morphometry meta-analysis of gray and white matter finds significant areas of differences in bipolar patients from healthy controls. *Bipolar Disord* 19(2):74–83. <https://doi.org/10.1111/bdi.12488>
79. Pinaya WHL, Gadelha A, Doyle OM et al (2016) Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Sci Rep* 6(1):38897. <https://doi.org/10.1038/srep38897>
80. Hibar DP, Westlye LT, Doan NT et al (2018) Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA bipolar disorder working group.

- Mol Psychiatry 23(4):932–942. <https://doi.org/10.1038/mp.2017.73>
81. Lin K et al (2018) Illness, at-risk and resilience neural markers of early-stage bipolar disorder. J Affect Disord 238:16–23
 82. Keenan KE, Gimbutas Z, Dienstfrey A et al (2021) Multi-site, multi-platform comparison of MRI T1 measurement using the system phantom. Lundberg P, ed. PLoS One 16(6): e0252966. <https://doi.org/10.1371/journal.pone.0252966>
 83. Yamada H, Abe O, Shizukuishi T et al (2014) Efficacy of distortion correction on diffusion imaging: comparison of FSL Eddy and Eddy_Correct using 30 and 60 directions diffusion encoding. Najbauer J, ed. PLoS One 9(11): e112411. <https://doi.org/10.1371/journal.pone.0112411>
 84. Fortin JP, Cullen N, Sheline YI et al (2018) Harmonization of cortical thickness measurements across scanners and sites. NeuroImage 167:104–120. <https://doi.org/10.1016/j.neuroimage.2017.11.024>
 85. Fortin JP, Parker D, Tunc B et al (2017) Harmonization of multi-site diffusion tensor imaging data. NeuroImage 161:149–170
 86. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8(1):118–127. <https://doi.org/10.1093/biostatistics/kxj037>
 87. Pomponio R, Erus G, Habes M et al (2020) Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage 208:116450.
 88. Eshaghzadeh Torbati M, Minhas DS, Ahmad G et al (2021) A multi-scanner neuroimaging data harmonization using RAVEL and ComBat. NeuroImage 245:118703. <https://doi.org/10.1016/j.neuroimage.2021.118703>
 89. Cetin-Karayumak S, Stegmayer K, Walther S et al (2020) Exploring the limits of ComBat method for multi-site diffusion MRI harmonization. Neuroscience. <https://doi.org/10.1101/2020.11.20.390120>
 90. Iwabuchi SJ, Liddle PF, Palaniyappan L (2013) Clinical utility of machine-learning approaches in schizophrenia: improving diagnostic confidence for translational neuroimaging. Front Psych 4:95. <https://doi.org/10.3389/fpsyg.2013.00095>
 91. Schulz MA, Bzdok D, Haufe S, Haynes JD, Ritter K (2022) Performance reserves in brain-imaging-based phenotype prediction. Neuroscience. <https://doi.org/10.1101/2022.02.23.481601>
 92. Antosik-Wójcińska AZ, Dominiak M, Chojnicka M et al (2020) Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling. Int J Med Inform 138:104131. <https://doi.org/10.1016/j.ijmedinf.2020.104131>
 93. Faerholt-Jepsen M, Bauer M, Kessing LV (2018) Smartphone-based objective monitoring in bipolar disorder: status and considerations. Int J Bipolar Disord 6(1):6. <https://doi.org/10.1186/s40345-017-0110-8>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Disclosure Statement of the Editor

Olivier Colliot is a civil servant of the French government, employed by the Centre National de la Recherche Scientifique. He reports that his work has been funded by the French government under management of the Centre National de la Recherche Scientifique, by the French government under management of Agence Nationale de la Recherche, by the European Union, by the Paris Brain Institute, by Inria, by the Abeona Foundation, and by the Fondation Vaincre Alzheimer. He reports having received consulting fees from AskBio and TheraPanacea. Members from his laboratory have co-supervised a PhD thesis with Qynapse. He reports that his spouse is an employee of myBrainTechnologies. He reports that he holds the following patent registered at the International Bureau of the World Intellectual Property Organization: PCT/IB2016/0526993, Schiratti J-B, Allassonniere S, Colliot O, Durrelman S, A method for determining the temporal progression of a biological phenomenon and associated methods and devices.

INDEX

A

- ABCD, *see* Adolescent brain cognitive development (ABCD)
Accelerometer 361–366,
 371–373, 760, 768, 793, 795, 864, 867
Adagrad 96–97
Adam 97–98, 700
ADC, *see* Apparent diffusion coefficient (ADC)
ADHD, *see* Attention deficit hyperactivity disorder (ADHD)
ADNI, *see* Alzheimer’s disease neuroimaging initiative (ADNI)
Adolescent brain cognitive development (ABCD) 680,
 681, 758, 771, 778, 980
AE, *see* Autoencoder (AE)
Affine registration 264, 406, 438–440
AIBL, *see* Australian imaging biomarkers and lifestyle study of aging (AIBL)
Akinesia 848
AlexNet 108, 553, 661, 662, 722
Alignment 135, 318, 323, 436,
 437, 440, 444, 447, 449, 454, 512, 784
ALS, *see* Amyotrophic lateral sclerosis (ALS)
Alzheimer 249, 372, 535, 758, 763, 772, 822
Alzheimer’s disease (AD) ix, 60, 234, 237,
 246, 247, 249, 254, 256, 265, 267, 273–276,
 302, 303, 326, 327, 370–372, 391, 462, 463,
 466, 468, 469, 472, 473, 476, 478, 491, 495,
 500, 503, 511, 514, 520–525, 534–536, 542,
 553, 554, 557, 559, 574, 577, 605, 619, 623,
 631–633, 657, 679–683, 685–688, 690–692,
 697, 700, 724, 726, 754, 758, 770, 773, 775,
 777, 791, 794, 807–833, 847, 848, 940, 1015
Alzheimer’s disease neuroimaging initiative (ADNI) 273, 376, 478,
 521, 523, 633, 679–681, 690, 692, 754, 758,
 770, 771, 773, 822, 826–828
Amyloid 273, 303, 463, 468, 472, 681, 687, 700,
 772, 773, 778, 808, 809, 812, 815–817, 822, 945
Amyotrophic lateral sclerosis (ALS) 247, 303,
 534, 760, 762, 763, 791, 811
Aneurysm 269, 701, 922, 937–940
Apparent diffusion coefficient (ADC) 267, 926,
 929–931, 968

- Area under the curve (AUC) 468–472,
 474, 604, 607, 608, 610, 611, 625, 731–734,
 741, 794, 827, 904, 905, 929, 932, 933, 935,
 936, 938–940, 971, 993, 998, 1016, 1021, 1024,
 1025, 1027
Arterial spin labeling (ASL) 758, 760,
 761, 922, 926, 932, 934, 935, 956
Arteriovenous malformation (AVM) 922, 923, 936
Artificial intelligence (AI) vii, 3–6, 8–11,
 21, 77, 225, 345, 347–349, 376, 476, 479, 480,
 534, 535, 537, 558–561, 717, 720, 739, 832,
 874, 937–940, 948, 963–973
ASL, *see* Arterial spin labeling (ASL)
Astrocytoma 247, 555, 729, 964, 965
Atlas 260, 264, 274, 320,
 326, 377, 404, 406, 412, 435, 436, 441, 453,
 499, 687, 829, 861, 881, 928, 930, 944, 948
Attention deficit hyperactivity disorder (ADHD) 371,
 500, 576, 758, 760, 763, 771, 781, 977–982,
 986, 987, 993–997, 1030
Attention U-Net 392–394, 554, 558, 560, 561, 932
AUC, *see* Area under the curve (AUC)
Australian imaging biomarkers and lifestyle study of aging (AIBL) 680, 681, 758, 770, 771, 773
Autism spectrum disorder (ASD) 12, 136,
 499, 500, 576, 645, 977–979, 981–984,
 986–989, 993–997, 1000, 1013, 1021, 1028
Autoencoder (AE) 13, 78,
 110–112, 134, 152–160, 176, 184, 185,
 211–212, 421–424, 445–447, 452, 471, 561,
 588–590, 659, 683, 684, 691, 825, 826, 929
AVM, *see* Arteriovenous malformation (AVM)
Axial 256–258, 266, 267, 427,
 428, 467, 848, 850, 861, 903, 931, 967

B

- Backpropagation 78, 85, 91–92,
 152, 168, 185, 588, 679–681
Batch normalization 95–96, 164, 170, 181, 427
BATS, *see* Brisbane adolescent twin study (BATS)
BCI, *see* Brain computer interface (BCI)
BD, *see* Bipolar disorder (BD)
Bias field correction 263–265, 404
Bioinformatics 319, 323–325, 719, 782

- Biomarker 256, 302, 303, 327, 368, 427, 454, 460, 462, 463, 466, 468, 472, 477, 495, 501, 502, 504, 513–517, 519, 520, 523–525, 536, 537, 540–542, 557, 561, 562, 575, 641, 681, 700, 758, 770, 773, 778, 786, 787, 808, 812, 815, 821–826, 828, 829, 849, 863, 869, 880, 889, 946, 968–969, 1014, 1015
- Bipolar disorder (BD) ix, 268, 371, 372, 534, 535, 758, 771, 994, 1009–1011, 1013–1016, 1018–1023, 1025–1031
- Blood-oxygen-level-dependent (BOLD) 267, 268, 769, 872, 904
- BOLD, *see* Blood-oxygen-level-dependent (BOLD)
- Brain computer interface (BCI) 303–307, 759, 760, 775
- Brain tumour segmentation challenge (BraTS) 176, 417, 426, 427, 470, 474, 475, 555, 969, 970, 972
- BraTS, *see* Brain tumour segmentation challenge (BraTS)
- Brisbane adolescent twin study (BATS) 761, 768, 793, 794
- C**
- CAD, *see* Computer assisted diagnosis (CAD)
- CADASIL, *see* Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL)
- Calibration 305, 306, 609, 610
- Canadian institute for advanced research (CIFAR) 86, 203, 207
- CBD, *see* Cortico-basal degeneration (CBD)
- CBF, *see* Cerebral blood flow (CBF)
- CBV, *see* Cerebral blood volume (CBV)
- Cerebellar variant of multiple system atrophy (MSA-C) 848, 853, 857, 861, 862
- Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) 810
- Cerebral blood flow (CBF) 267, 270, 275, 923, 925, 926, 932, 934–935
- Cerebral blood volume (CBV) 270, 923, 925, 926, 968
- Cerebral microbleed (CMB) 940, 941, 945–946, 948
- Cerebral small vessel disease (cSVD) 922, 940–949
- Cerebrospinal fluid (CSF) 264, 327, 405, 460, 463, 468, 472, 682, 700, 758, 815, 824, 829, 853, 857, 860, 863, 869, 909, 936, 947, 1030
- Cerebrovascular ix, 270, 374, 921–949
- CIFAR, *see* Canadian institute for advanced research (CIFAR)
- Classification viii, 10, 25, 79, 118, 140, 203, 249, 268, 286, 338, 356, 399, 446, 460, 498, 534, 602, 657, 707, 774, 823, 849, 927, 965, 983, 1010
- Clinically isolated syndrome (CIS) 902, 912, 913
- Clustering 12, 13, 15, 26, 60–66, 301, 303, 492–505, 517, 524, 525, 527, 551, 555, 557, 764, 828, 868, 881, 888, 889, 906, 931, 987–989, 992–994, 996, 1015, 1019–1020, 1023, 1027
- CMB, *see* Cerebral microbleed (CMB)
- CNN, *see* Convolutional neural network (CNN)
- Coding schemes 339, 340, 344
- Cognitive 4, 234, 256, 288, 358, 460, 511, 612, 681, 758, 808, 848, 879, 899, 940, 965, 980, 1010
- Computational pathology viii, 533–563
- Computed tomography (CT) vii, 144, 145, 149, 173, 176, 177, 183, 253, 254, 256, 259, 269–271, 274, 277, 392, 396, 425, 447, 449, 465, 471, 553, 660, 679–681, 686, 689, 701, 707, 721, 725, 726, 732, 766, 768, 815, 831, 922–925, 927, 929–932, 934–938, 940
- Computer assisted diagnosis (CAD) 937
- Confidence interval viii, 602, 620–622, 624, 625, 633, 641, 740, 744, 746
- Connected objects vii, 355–383
- Convolution 10, 78, 133, 142, 203, 392, 446, 552, 662, 928, 997
- Convolutional neural network (CNN) vii, 77–112, 133–136, 140, 142, 153, 161, 174, 175, 182, 185, 203, 205–207, 210, 212, 213, 216, 222, 224, 392, 394, 408, 410, 411, 414, 437, 446–450, 468, 474, 551–553, 555–557, 561, 563, 632, 660, 661, 666, 676, 681, 682, 684–686, 688, 689, 698, 722, 853–855, 857, 886, 904, 913, 927, 930–932, 934, 935, 938, 939, 944–946, 948, 969, 971
- Coronal 256, 257, 428, 688, 883, 903, 931
- Cortico-basal degeneration (CBD) 557, 849, 850
- Cost function 17, 19, 21, 31, 37, 39, 41, 71, 72, 87–90, 110, 164, 377, 437, 438, 441–445
- Cross-attention 196–197, 199, 201, 202, 209, 212, 214, 216, 219, 220, 225, 411
- Cross-entropy loss 50, 88, 164, 172, 206, 215, 398, 399, 497
- Cross-sectional 495, 512–518, 523, 526, 758–766, 768, 786, 787, 828, 831, 868, 906–908
- Cross-validation (CV) 427, 469, 501, 615–619, 622, 624, 626, 713, 714, 846, 858–862, 866, 867, 869, 870, 872, 873, 886, 891, 909, 928, 933, 985, 995, 998, 1022, 1023, 1025–1027

- cSVD, *see* Cerebral small vessel disease (cSVD)
 CT, *see* Computed tomography (CT)
 CV, *see* Cross-validation (CV)
 CycleGAN 176–178, 404, 448
- D**
- Data augmentation 95, 376, 408–409, 417, 427, 428, 535, 747
 Database of genotypes and phenotypes
 (dbGAP) 765, 783, 784, 787, 788
 Data-driven disease progression modelling
 (D3PM) 468, 511–527, 828, 869
 Data harmonization 404, 478, 823, 828–830
 Data leakage 602, 617–619, 633, 640, 641, 649, 829
 Data split 426, 615, 619, 633, 638, 640–642
 Data warehouse 332, 766, 789–790
 dbGAP, *see* Database of genotypes and phenotypes
 (dbGAP)
 Decoder 110, 111, 127–129, 134, 135, 153, 154, 156, 183, 185, 186, 194, 199–202, 209, 211, 212, 215, 392, 409–412, 421, 423, 424, 427, 448, 676, 930, 932, 938, 939
 Deep learning (DL) vii, 6, 10, 11, 13, 18, 73, 77–112, 166, 198, 202, 214, 224, 258, 261, 264, 277, 348, 360, 391–428, 435–454, 460, 469, 470, 474, 497, 501, 534, 537, 544, 549, 552–554, 557, 559, 561, 563, 618, 619, 638, 647, 657, 659, 665, 669, 675, 681, 707, 719, 722, 727, 729, 789, 824–827, 859, 865, 882, 883, 885, 886, 890, 905, 909, 927, 931, 937, 939, 942–946, 963, 969–972, 983, 984, 986, 991–993, 997–998, 1014
 Deformable registration 406, 438, 440, 441, 445
 Dementia ix, 234, 254, 303, 459, 499, 511, 535, 681, 768, 807, 848, 922
 Dementia with Lewy bodies (DLB) ix, 276, 463, 469, 811–812, 821, 848, 850
 Diagnosis 12, 48, 136, 234, 253, 303, 342, 371, 459, 500, 534, 576, 642, 690, 705, 768, 808, 849, 902, 964, 979, 1010
 DIAN, *see* Dominantly inherited Alzheimer network (DIAN)
 Dice loss 398–400, 427
 Dice similarity coefficient (DSC) 401, 402, 425, 553, 728, 926, 930–932, 936, 939, 943, 944, 947
 Differential diagnosis 48, 234, 240, 461–464, 467–469, 471, 476, 480, 812, 817, 823, 825, 827, 833, 849, 851, 853, 861–862, 873, 903–905, 915
 Diffusion models vii, 148, 150, 186–187, 215
 Diffusion tensor imaging (DTI) 267, 468, 573, 829, 861, 862, 887, 903, 904, 909
 Diffusion-weighted Imaging (DWI) 469, 471, 574, 755, 758–762, 769, 820, 861, 902, 903, 922, 926, 929–933, 941, 985, 998, 1017–1020, 1030
 Digital pathology 534, 536, 537, 544–545, 552
 Dimensionality reduction 12, 13, 26, 63, 66–70, 494, 555, 618, 619, 682, 683, 717, 871
 Direct problem 298
 DL, *see* Deep learning (DL)
 DLB, *see* Dementia with Lewy bodies (DLB)
 Domain adaptation 178, 188, 425, 447–449
 Dominantly inherited Alzheimer network
 (DIAN) 822, 828
 Dopamine 238, 247, 275, 689, 701, 811, 848, 850, 857, 860, 863, 869, 871–872
 D3PM, *see* Data-driven disease progression modelling
 (D3PM)
 DSC, *see* Dice similarity coefficient (DSC)
 DTI, *see* Diffusion tensor imaging (DTI)
 DWI, *see* Diffusion-weighted Imaging (DWI)
- E**
- EBM, *see* Event-based model (EBM)
 EEG, *see* Electroencephalography (EEG)
 EfficientNet 109–110, 556
 EHR, *see* Electronic health record (EHR)
 Electrode 99, 247, 289, 291, 292, 296, 300, 681, 682, 774, 775, 872, 884
 Electroencephalography (EEG) vii, 248, 285–307, 356, 380, 460, 481, 500, 576, 679, 681, 682, 755, 757–760, 762, 764, 766, 768, 774–775, 780, 830, 855–857, 872, 884, 986, 1013, 1016, 1017, 1023, 1024
 Electronic health record (EHR) vii, 331–351, 594, 714, 755, 757, 761, 789–792, 995
 Encoder 110, 111, 127–129, 134, 135, 153–156, 188, 194, 199–203, 209, 211, 212, 214–216, 220–221, 223, 392, 394, 395, 409–412, 421, 423, 424, 448, 675, 691, 693, 930, 932, 938, 939
 Enhancing neuroimaging genetics through meta-analysis
 (ENIGMA) 758, 771, 772, 780–781, 1020, 1027, 1030
 ENIGMA, *see* Enhancing neuroimaging genetics through meta-analysis (ENIGMA)
 Epilepsy ix, 246–248, 273, 302–303, 327, 370, 373, 480, 536, 576, 726, 760, 773–775, 792, 879–891, 978, 996
 Event-based model (EBM) 495, 512, 514–517, 523, 524, 828, 855, 869, 912
 Evoked activity 287–288, 299

F

FCD, *see* Focal cortical dysplasia (FCD)

Feature extraction 172, 299–302, 366, 377, 403, 446–447, 639, 647, 682, 698, 707, 722, 754, 860, 928, 991

Features 5, 26, 77, 118, 164, 199, 236, 254, 285, 319, 332, 358, 392, 435, 464, 493, 512, 534, 575, 614, 639, 655, 707, 754, 808, 848, 880, 899, 927, 963, 978, 1011

Feature selection 32, 34, 299–302, 366, 618, 619, 657, 721, 722, 858, 860, 866, 936, 988, 1018, 1021, 1025

FLAIR, *see* Fluid attenuated inversion recovery (FLAIR)

Fluid attenuated inversion recovery (FLAIR) 263, 265, 412, 426, 427, 469, 471, 555, 755, 758, 760, 761, 769, 818, 820, 825, 882–886, 890, 902, 903, 906, 907, 909, 911, 922, 926, 929, 930, 941, 943, 944, 946, 969, 970, 972

fMRI, *see* Functional magnetic resonance imaging (fMRI)

Focal cortical dysplasia (FCD) 879, 884–888

Focal loss 399

Fronto-temporal lobar degeneration
(FTLD) 463, 468, 469

Functional magnetic resonance imaging (fMRI) 267, 268, 358, 502, 573, 574, 576, 581, 680, 681, 755, 759, 769, 821, 853, 856, 857, 993, 994, 1013, 1017–1025, 1028, 1030, 1031

G

GAN, *see* Generative adversarial network (GAN)

Gated recurrent unit (GRU) 118, 124–125, 932, 944

Gaussian mixture model (GMM) 26, 60, 63–66, 147–149, 424, 497, 824, 825, 857, 859, 942, 944, 946, 948, 988

Gaussian process 512, 520, 521, 523, 689, 826, 990, 1016, 1028

GBM, *see* Glioblastoma (GBM)

GDL, *see* Generalized Dice loss (GDL)

Generalization 78, 92–96, 400, 495, 496, 550, 574, 602, 615, 617, 619, 620, 622–625, 825, 984, 1000, 1021

Generalized Dice loss (GDL) 398–400

Generative adversarial network (GAN) vii, 112, 139–188, 209, 212, 404, 424–425, 448, 496, 497, 534, 550, 659, 909, 928

Generative models vii, 112, 134, 139–188, 212, 376, 496, 524, 577, 579, 591, 659

Genetic FTD initiative (GENFI) 822, 828

GENFO, *see* Genetic FTD initiative (GENFI)

Genome-wide association study (GWAS) 326, 755, 764, 783, 789, 813, 989, 994

Genomics 315–320, 322,

324–327, 474, 504, 593, 756–758, 760, 762–764, 766, 768, 770, 780–783, 785–789, 887, 889, 989, 997, 998, 1030

Glioblastoma (GBM) 247, 467, 471, 476, 504, 964–969, 971, 972

Glioma 176, 275, 463, 464, 467, 469–471, 474, 552, 553, 555–557, 963–972

Global positioning system (GPS) 367, 371, 372, 378, 793

GMM, *see* Gaussian mixture model (GMM)

GPS, *see* Global positioning system (GPS)

GPU, *see* Graphical processing unit (GPU)

Grad-CAM 551, 561, 660, 666, 667, 670, 679–681, 685, 686, 694–696, 698

Gradient descent 19, 20, 35, 78, 79, 89–91, 93, 96, 132, 153, 588, 700

Graph-based 259, 260, 265, 267, 274, 276

Graphical processing unit (GPU) 10, 78, 91, 180, 183, 408, 638, 640, 790

Gray matter (GM) 264, 265, 405, 468, 502, 554, 684, 688, 690, 769, 817–821, 824, 860, 884, 889, 912–914, 979, 1016, 1018, 1027–1029

GRU, *see* Gated recurrent unit (GRU)

GWAS, *see* Genome-wide association study (GWAS)

Gyroscope 361–366, 795, 858, 864, 865, 867, 872

H

HAR, *see* Human activity recognition (HAR)

HGG, *see* High-grade glioma (HGG)

Hierarchical clustering 492–494, 500, 502, 503, 993

High-grade glioma (HGG) 467, 469, 553, 972

High-throughput 315, 317–319, 321, 491, 785

Hippocampal sclerosis (HS) 809, 880–883

Hippocampus 264, 425, 473, 631, 686–688, 808, 811, 824–826, 866, 880, 881, 996, 1018

Histology 474, 504, 536, 540, 553, 555, 562, 563, 681, 888, 964

Histopathology 109, 175, 176, 474, 540, 541, 543, 549, 551, 558, 560–563, 884, 889

HS, *see* Hippocampal sclerosis (HS)

Human activity recognition (HAR) 356, 360–367, 375

I

IA, *see* Intracranial aneurysm (IA)

ID, *see* Intelligence disabilities (ID)

Idiopathic rapid eye movement sleep behavior disorder (iRBD) 852, 857, 859

Inertial systems 357, 360–362

Intelligence disabilities (ID) 341, 786, 788, 977, 979, 981

Intensity rescaling 263–264
 Interpretability viii, 38, 333, 477,
 479, 480, 559, 574, 575, 592, 655–701, 744,
 988, 992, 999–1000
 Intracranial aneurysm (IA) 922, 923,
 935, 937–940
 Inverse problem 298–299
iRBD, *see* Idiopathic rapid eye movement sleep behavior disorder (iRBD)
 The ischemic stroke lesion segmentation (ISLES) 417, 930, 931, 933

K

KDE, *see* Kernel density estimate (KDE)
 Kernel 9, 26, 39, 41, 42, 44–47,
 70–73, 100–106, 108, 393, 406, 409, 472, 512,
 582–583, 591, 661, 662, 864, 1020, 1026
 Kernel density estimate (KDE) 405–406,
 512, 515, 527
k-means 13, 26, 60–63, 65,
 66, 69, 492–494, 500–502, 524, 551, 555, 855,
 906, 931, 988, 1020, 1023, 1027
k-nearest neighbors (kNN) 28, 336, 721,
 851, 852, 854, 855, 857, 859, 860, 864, 866, 942
 kNN, *see* *k*-nearest neighbors (kNN)
 Kullback–Leibler divergence (KL/KLD) 111, 154,
 156–160, 162, 163, 166, 416, 423, 589

L

Lacunes 809, 810, 819,
 825, 940, 941, 946–948
 Large vessel occlusion (LVO) 464, 737, 923–929
 LASSO, *see* Least absolute shrinkage and selection operator (LASSO)
 LATE, *see* Limbic-predominant age-related TDP-43 encephalopathy (LATE)
 Latent variable 86, 89, 134,
 146, 154, 186, 208, 575, 577–591, 691, 868
 Layer-wise relevance (LRP) 660, 668,
 669, 679–681, 685, 686, 693–695, 697, 698
 LDA, *see* Linear discriminant analysis (LDA)
 Least absolute shrinkage and selection operator (LASSO) 657, 904, 1021, 1024
 Lesion viii, 11, 235, 265, 374, 391, 462,
 543, 576, 686, 711, 769, 809, 880, 899, 922, 969
 LGG, *see* Low-grade glioma (LGG)
 Limbic-predominant age-related TDP-43 encephalopathy (LATE) 809
 Linear discriminant analysis (LDA) 26, 55, 69–71,
 852–854, 856, 857, 883, 1023
 Linear regression 9, 14, 19, 20,
 25, 26, 30–32, 35, 38, 40, 42, 523, 578, 579, 583,
 585, 657, 829, 865, 944, 1021

Logistic regression (LR) 4, 13, 25, 26, 32–35,
 37, 42, 48–51, 86, 258, 471, 472, 555, 607, 621,
 852–858, 860, 862, 863, 866, 873, 904, 929,
 986, 993, 997, 998, 1021–1025
 Logopenic progressive aphasia (LPA) 809
 Longitudinal vii, 136, 262, 336, 373, 392, 414,
 440, 473, 503, 512–514, 519–525, 619, 692,
 757–768, 770, 771, 777, 786, 791, 828, 829,
 831, 868–870, 881, 906, 908, 941, 946, 1016
 Long short-term memory (LSTM) 118, 122–125,
 133–136, 187, 194, 202, 676, 852, 857, 859,
 873, 933
 Loss 16, 27, 79, 119, 139, 205, 265, 318, 333,
 367, 397, 445, 465, 497, 511, 581, 639, 664,
 718, 808, 848, 880, 899, 923, 965, 997, 1010
 Low-grade glioma (LGG) 467, 469, 474, 553, 972
 LPA, *see* Logopenic progressive aphasia (LPA)
 LR, *see* Logistic regression (LR)
 LRP, *see* Layer-wise relevance (LRP)
 LSTM, *see* Long short-term memory (LSTM)
 LVO, *see* Large vessel occlusion (LVO)

M

Machine learning (ML) 3, 25, 77,
 348, 355, 437, 460, 491, 511, 536, 575, 601,
 631, 655, 705, 755, 789, 822, 847, 880, 905,
 949, 963, 979, 1009
 Magnetic resonance imaging (MRI) vii, 19,
 144–149, 176, 254, 256–270, 274, 277, 298,
 396, 404, 405, 412, 413, 425, 426, 447, 464,
 468, 472, 474, 477, 479, 555–558, 561, 590,
 621, 632, 645, 660, 681, 686, 700, 707, 726,
 754, 756–769, 809, 880, 922, 924, 1028
 Magnetoencephalography (MEG) viii, 285–307,
 460, 755, 757–762, 766, 768, 774, 775
 Major depressive disorder (MDD) 500, 502,
 758, 760, 763, 771, 853, 857, 1010, 1013, 1014,
 1016, 1018, 1020–1028
 Major depressive episode (MDE) 1010
 Malignant 248, 464, 723,
 745, 767, 787, 791, 964
 Markov chain monte carlo (MCMC) 150, 515
 MDD, *see* Major depressive disorder (MDD)
 MDE, *see* Major depressive episode (MDE)
 Mean squared error (MSE) 48, 58, 88,
 107, 154, 416, 422, 441, 449, 700, 718, 719
 MEDA, *see* Minimal evidence of disease activity (MEDA)
 Medical segmentation decathlon (MSD) 425
 MEG, *see* Magnetoencephalography (MEG)
 Meningioma 247, 471, 729, 965
 Messenger RNA (mRNA) 315, 321, 324, 325
 Metabolomics vii, 315, 317, 318,
 320, 323, 324, 327, 758, 761–763, 998, 1013

- MI, *see* Mutual information (MI)
- Microbleed ix, 818, 921, 935, 940, 945, 947
- Minimal evidence of disease activity (MEDA) 907
- ML, *see* Machine learning (ML)
- MLP, *see* Multi-layer perceptron (MLP)
- MND, *see* Motor neuron disease (MND)
- Mobile devices vii, 355–383
- Morphometry 453, 454, 468, 552, 860, 889, 999, 1019, 1025, 1028
- Motor 235, 237, 238, 249, 250, 264, 276, 288, 304, 305, 371, 373, 382, 701, 759, 774–776, 793, 810–812, 848–851, 862–865, 867–869, 871, 872, 889, 891, 899, 965, 977
- Motor neuron disease (MND) 371, 768, 776, 811
- MRI, *see* Magnetic resonance imaging (MRI)
- mRNA, *see* Messenger RNA (mRNA)
- MS, *see* Multiple sclerosis (MS)
- MSA, *see* Multiple system atrophy (MSA)
- MSD, *see* Medical segmentation decathlon (MSD)
- MSE, *see* Mean squared error (MSE)
- Multi-layer perceptron (MLP) 81–83, 102, 104, 126, 199–204, 217–219, 222, 398, 852, 854, 857
- Multimodal viii, 196, 197, 212–216, 225, 256, 372, 392, 413, 425, 444, 468, 475, 550, 562, 573–594, 769, 770, 823, 830, 882, 889, 890, 969, 986, 995, 1030, 1031
- Multiple sclerosis (MS) ix, 247, 249, 263, 265, 274, 324, 342, 343, 370, 392, 410, 411, 414, 417, 425, 427, 428, 462, 465, 466, 473, 474, 476, 536, 680, 686, 701, 726, 768, 769, 773, 792, 899–915
- Multiple system atrophy (MSA) ix, 848, 850, 857
- Mutual information (MI) 443–444
- N**
- National Institutes of Health (NIH) 332, 771, 777, 783, 787, 927
- Natural language processing (NLP) vii, 15, 136, 193, 199, 202, 203, 225, 236, 337, 338, 346, 360
- NAWM, *see* Normal appearing white matter (NAWM)
- NCC, *see* Normalized cross correlation (NCC)
- NDDs, *see* Neurodevelopmental disorders (NDDs)
- Neural network (NN) 8–11, 78, 82–95, 99, 102, 106, 109, 110, 118, 130, 132, 142, 154, 155, 169, 177, 186, 270, 377, 412, 421, 423, 446, 468, 551, 558, 588, 590, 591, 593, 594, 620, 659, 668, 699, 722, 824, 873, 883, 890, 928, 931, 938, 971, 986, 991
- Neurodevelopmental disorders (NDDs) 499, 701, 977–1000
- Neuroimaging vii, 240, 253–277, 371, 376, 377, 380, 391, 392, 403, 412, 460, 462, 463, 478, 491–505, 521, 575–577, 581, 592, 593, 648, 655–701, 754, 769–774, 780, 822, 879–891, 940, 946, 981, 988, 989, 991–993, 998, 1013, 1016, 1018, 1019, 1030
- Neurology vii, 241, 246, 249, 250, 275, 307, 512, 525, 691, 761, 763, 1017
- Neuromyelitis optica spectrum disorder (NMOSD) 904, 905
- Neuropsychology 240, 241
- NIH, *see* National Institutes of Health (NIH)
- NLP, *see* Natural language processing (NLP)
- NMOSD, *see* Neuromyelitis optica spectrum disorder (NMOSD)
- NN, *see* Neural network (NN)
- Normal appearing white matter (NAWM) 405, 914
- Normalized cross correlation (NCC) 442, 447
- O**
- Obsessive-compulsive disorder (OCD) 500, 758, 771, 978, 994, 1009, 1010
- OCD, *see* Obsessive-compulsive disorder (OCD)
- Oligodendrogioma 247, 555, 965
- Optimization 4, 9, 17, 31, 35, 38–42, 78, 88–98, 126, 132, 146, 159, 160, 163, 172, 323, 358, 419, 438, 441, 442, 444, 445, 452, 467, 480, 498, 499, 548, 553, 578, 580, 581, 583–585, 588–593, 620, 621, 663, 672, 711, 719, 722, 906, 939
- Oscillatory activity 287–289, 299, 302, 303
- Overfitting 17, 25, 32, 34–37, 40, 58, 78, 92–96, 445, 556, 642, 714, 723, 944, 984–986, 988, 990, 993, 998
- P**
- Parkinson ix, 144, 235, 238, 246, 247, 249, 276, 370, 372, 374, 391, 478, 514, 516, 534–536, 577, 645, 681, 689, 701, 772, 781, 791, 794, 795, 812, 847–874
- Parkinson variant of multiple system atrophy (MSA-P) 848, 853, 857, 861, 862
- Parkinson's disease (PD) ix, 144, 238, 246, 247, 249, 276, 370–374, 391, 478, 514, 516, 534–536, 577, 645, 681, 689, 701, 772, 781, 791, 794, 795, 812, 847–874
- Parkinson's progression markers initiative (PPMI) 276, 478, 680, 681, 772, 852, 856, 860, 869

- Partial least squares (PLS) 472, 575, 579–585, 588
- Partial volume 936, 944
- Patch 185, 188, 203, 204, 206–208, 210, 215, 217, 220–222, 408–410, 446, 447, 449, 549–553, 555, 556, 558–561, 577, 670, 672, 688, 690, 691, 698, 824, 826, 882, 885, 886, 928, 930, 938, 944
- PCA, *see* Posterior cortical atrophy (PCA); Principal component analysis (PCA)
- PD, *see* Parkinson’s disease (PD)
- Performance assessment viii, 171, 705–747
- Performance metrics viii, 171, 478, 479, 602–613, 625, 712, 718, 730, 732, 733, 736, 738, 741, 745, 746
- Perivascular spaces (PVS) 819, 825, 940, 941, 946–948
- PET, *see* Positron emission tomography (PET)
- p-hacking 633, 649
- Pick’s disease (PiD) 557
- PiD, *see* Pick’s disease (PiD)
- PLS, *see* Partial least squares (PLS)
- PML, *see* Progressive multifocal leukoencephalopathy (PML)
- Positioning systems 357, 360–362, 367
- Positron emission tomography (PET) vii, 165, 253, 254, 256, 257, 259, 271–277, 392, 463, 468, 472, 473, 558, 573, 574, 581, 619, 700, 707, 726, 755, 757, 758, 772–774, 778, 811, 815–817, 819, 822, 826–828, 860, 909, 910, 934, 935
- Posterior cortical atrophy (PCA) 809, 828
- PPA, *see* Primary progressive aphasia (PPA)
- PPMI, *see* Parkinson’s progression markers initiative (PPMI)
- PPMS, *see* Primary progressive multiple sclerosis (PPMS)
- Prediction 9, 27, 79, 118, 151, 206, 265, 401, 446, 459, 521, 551, 577, 601, 657, 717, 775, 823, 849, 880, 905, 923, 963, 980, 1014
- Preregistration 633, 642, 1000
- Primary progressive aphasia (PPA) 246, 811
- Primary progressive multiple sclerosis (PPMS) 343, 900, 912
- Principal component analysis (PCA) 26, 66–67, 300, 366, 551, 555, 582
- Progressive multifocal leukoencephalopathy (PML) 901–903
- Progressive supranuclear palsy (PSP) ix, 286, 287, 557, 811, 848, 850
- Proteomics vii, 315–318, 320, 323–327, 757, 758
- PSP, *see* Progressive supranuclear palsy (PSP)
- Psychiatry vii, 376, 758, 777, 781, 985, 989, 1012–1017, 1031
- Q**
- QTAB, *see* Queensland twin adolescent brain (QTAB)
- QTIM, *see* Queensland twin imaging (QTIM)
- Queensland twin adolescent brain (QTAB) 760, 772, 776, 793, 794
- Queensland twin imaging (QTIM) 761, 772, 776, 777
- R**
- Random forest 13, 26, 58–59, 460, 471–473, 722, 826, 858, 860, 862, 865, 866, 869, 873, 929, 930, 932, 940, 945, 947, 1019, 1022, 1025
- Rapid eye movement (REM) 811, 859, 870
- Reader study 707, 726, 740–744, 747
- Rectified linear unit (ReLU) 85–87, 91, 100, 101, 107, 108, 392, 393, 395, 665–668, 695, 698
- Recurrent neural networks (RNNs) vii, 78, 117–137, 149, 194, 202, 460, 551, 691, 692, 826, 856, 857, 859, 870
- Registration viii, 263–265, 403, 406, 427, 435–454, 558, 642, 679, 708, 777, 794, 824, 907, 928, 1000
- Regression 9, 25, 88, 151, 258, 336, 446, 460, 513, 578, 602, 657, 851, 904, 929, 983, 1019
- Regulatory viii, 250, 376, 378, 381, 517, 549, 614, 637, 649, 705–747, 808, 997
- Reinforcement learning 12–14, 551, 872
- Relapsing remitting multiple sclerosis (RRMS) 343, 900, 901, 905, 912, 913
- Reliability 243, 246, 301, 305, 479, 635, 659, 661, 677, 687, 696, 777, 865
- ReLU, *see* Rectified linear unit (ReLU)
- REM, *see* Rapid eye movement (REM)
- Repeatability 534, 632, 635
- Replicability 632, 635
- Representativeness 350, 715–717
- Reproducibility viii, 319, 382, 475, 479, 501, 549, 631–649, 690, 719, 720, 728, 747, 753, 786, 827, 830, 968, 969, 989, 993, 999, 1030
- Residual neural network (ResNet) 109, 140, 215, 448, 553, 557, 561, 682, 935, 938
- Ribonucleic acid (RNA) 313–316, 318, 319, 323–326, 563, 766, 787, 788, 821, 968
- Ridge regression 26, 39, 41, 72
- Rigidity 237, 238, 848
- Rigid registration 264, 439, 440
- RMSE, *see* Root mean square error (RMSE)
- RNA, *see* Ribonucleic acid (RNA)
- RNNs, *see* Recurrent neural networks (RNNs)
- Root mean square error (RMSE) 612, 613

S

- Sagittal 256, 257, 428, 886, 903, 931
 Saliency 659, 665, 696
 Schizophrenia ix, 246, 268, 371, 491, 495, 501, 534, 535, 726, 763, 771, 979, 994, 996, 1009, 1011–1017, 1019, 1028, 1030, 1031
 Secondary progressive multiple sclerosis (SPMS) 343, 900
 Segmentation 13, 88, 140, 207, 263, 363, 391, 435, 470, 534, 602, 645, 682, 707, 824, 881, 906, 923, 966
 Self-attention 128, 196–199, 201–203, 208, 209, 218, 219
 Semi-supervised learning (SSL) 415–417, 419
 Sensors vii, 289–292, 296–298, 300, 355–383, 460, 757–760, 762, 764, 768, 774, 775, 793–795, 851, 858, 861, 864, 865, 867, 872, 874
 Similarity function 437, 447
 Single-cell 316, 323–324, 326, 757, 787, 821, 968
 Single-nucleotide polymorphism (SNP) 324, 326, 574, 592–594, 761, 780, 784, 979, 997, 1014
 Single-photon emission computed tomography (SPECT) vii, 253, 254, 256, 259, 275–277, 680, 681, 689, 701, 707, 811
 Skull stripping 264, 403, 404, 427
 Smartphone 349, 356, 360, 361, 366, 367, 369–371, 373–375, 460, 756, 758, 760, 762, 764, 766, 768, 793–795, 832, 851, 858, 1031
 SNP, *see* Single-nucleotide polymorphism (SNP)
 Source reconstruction 295, 297
 Sparsity 39, 111, 498, 593
 Spatial normalization 274, 276, 406, 453, 829
 SPECT, *see* Single-photon emission computed tomography (SPECT)
 SPMS, *see* Secondary progressive multiple sclerosis (SPMS)
 SSD, *see* Sum of squared difference (SSD)
 SSL, *see* Semi-supervised learning (SSL)
 Standardized uptake value ratio (SUVR) 274, 817, 862
 Statistical testing 619–622, 624
 STR, *see* Swedish twin registry (STR)
 Stratification 492, 500, 514, 552, 557, 559, 576, 613–617, 684, 685, 717, 886–888, 913–915, 983, 987, 988, 1025
 Stroke 143, 237, 270, 303, 333, 392, 459, 536, 576, 737, 792, 810, 879, 921
 StyleGAN 178–180
 Substance use disorder (SUD) 535, 758, 763, 1009

- Subtyping 267, 277, 464, 491–495, 497, 499–501, 503–505, 512, 513, 517, 524, 526, 527, 993, 1016
 SUD, *see* Substance use disorder (SUD)
 Sum of squared difference (SSD) 441, 442
 Supervised learning 12–16, 20, 25, 26, 60, 110, 376, 406, 415, 420, 985
 Support vector machine (SVM) 9, 13, 25, 26, 42–48, 460, 468, 471, 472, 474, 496, 498, 553, 556, 621, 721, 852–861, 863–866, 882, 883, 905, 913, 919, 929, 933, 946, 969, 970, 986, 1013, 1028
 SUVR, *see* Standardized uptake value ratio (SUVR)
 SVM, *see* Support vector machine (SVM)
 Swedish twin registry (STR) 761, 776
 Synucleinopathy 848

T

- Tau 273, 463, 472, 473, 542, 553, 554, 557, 558, 700, 808, 810, 812, 815–817, 848, 850
 Temporal lobe epilepsy (TLE) 880–883, 887–891
 Testing set/test set 92, 93, 400, 612, 614–622, 624, 625, 633, 641, 692, 696, 714, 715, 718, 859, 864, 867, 935, 936, 939, 949, 985
 TLE, *see* Temporal lobe epilepsy (TLE)
 Training set 12–14, 16, 17, 36, 79, 87–89, 91, 92, 111, 399, 400, 426, 591, 614, 615, 618, 619, 633, 687, 714, 715, 718, 832, 883, 943, 987
 Transcriptomics vii, 315–319, 322, 324, 325, 327, 504, 549, 558, 563, 758, 762, 763, 998
 Transformation model 437–441, 446, 449, 452
 Transformer vii, 112, 128–129, 140, 184, 185, 187, 188, 193–225, 396–398, 411, 449–451
 Tremor 238, 239, 246, 247, 276, 373, 701, 847, 848, 854, 858, 865, 868, 872, 873
 Tumor 11, 104, 176, 248, 254, 326, 394, 459, 463, 500, 533, 603, 644, 701, 707, 773, 815, 879, 963
 Tversky loss 400

U

- UKB, *see* UK biobank (UKB)
 UK biobank (UKB) 754, 757, 761, 769, 771, 778–781, 784, 793, 794, 822, 1030
 Underfitting 35, 36, 39, 94
 U-Net 111, 151, 170, 171, 173–176, 392–398, 410–412, 427, 447, 449, 451, 551, 553, 554, 556, 558, 560, 561, 824, 825, 905, 927, 928, 930–935, 938, 943–945
 United States food and drug administration (FDA) 534, 705–708
 Unsupervised learning 12–14, 26, 60, 110, 415, 445, 512, 534, 551, 987–989

V

- VaD, *see* Vascular dementia (VaD)
 VAE, *see* Variational autoencoder (VAE)
 Validation 92, 171, 250, 305,
 333, 377, 400, 468, 501, 552, 601, 633, 713,
 824, 859, 882, 914, 927, 984, 1030
 Validation set 92, 93, 400,
 469, 617–619, 622, 623, 641, 714
 Variational autoencoder (VAE) 13, 134, 152–160,
 184–186, 412, 423–424, 452, 588, 589, 659
 Vascular cognitive impairment (VCI) 463, 810
 Vascular dementia (VaD) ix, 809–810,
 812, 821, 940
 VCI, *see* Vascular cognitive impairment (VCI)
 Vertex-based 258, 259,
 265, 268, 274, 276, 1019
 VETSA, *see* Vietnam era twin study of aging (VETSA)
 Vietnam era twin study of aging
 (VETSA) 761, 772, 776, 777
 V-Net 392, 393, 936
 Voxel-based 258, 259, 264, 265, 267, 268,
 270, 274, 276, 453, 468, 860, 884, 948, 1019

W

- Wasserstein 166–170, 172
 Wearables 291, 356, 357,
 367, 370, 373, 374, 378–380, 383, 463, 757,
 770, 793, 821, 832, 851, 865
 WGS, *see* Whole genome sequence (WGS)
 White matter hyperintensity (WMH) 941–943
 White matter (WM) viii, 264, 265,
 267, 270, 273, 343, 405, 541, 576, 682, 684,
 700, 701, 769, 777, 809, 817–820, 824, 825,
 860, 869, 884, 909, 913, 914, 929, 940–943,
 996, 1015, 1017, 1019, 1025, 1028–1030
 WHO, *see* World Health Organization (WHO)
 Whole genome sequence (WGS) 758, 760,
 779, 780, 783–785, 787, 788
 Whole slide image (WSI) 533–535, 537,
 541, 542, 544–559, 561, 563
 WM, *see* White matter (WM)
 WMH, *see* White matter hyperintensity (WMH)
 World Health Organization (WHO) 326, 341,
 342, 464, 967
 WSI, *see* Whole slide image (WSI)