HELLENIC DEMOCRACY
UNIVERSITY OF WESTERN MACEDONIA
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL
& COMPUTER ENGINEERING

AI Alzheimer And Dementia Classification

A Review of Neuroimaging and Deep Learning Approaches

Gavriilidis Paraskevas

Supervisor: Prof. Fragulis Georgios

A thesis submitted in partial fulfillment for the

degree of MSc in Electrical Engineering

Department of Electrical and Computer Engineering
University of Western Macedonia

KOZANI / January / 2026

HELLENIC DEMOCRACY
UNIVERSITY OF WESTERN MACEDONIA
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL
& COMPUTER ENGINEERING

DECLARATION OF ORIGINALITY AND ASSUMPTION OF PERSONAL RESPONSIBILITY

I hereby declare that, in accordance with Article 8 of Law 1599/1986 and Articles 2, 4, 6 par. 3 of Law 1256/1982, this thesis entitled:

"AI Alzheimer And Dementia Classification"

as well as the electronic files and source code developed or modified within the framework of this work and explicitly mentioned in the accompanying text, which has been prepared at the Department of Electrical and Computer Engineering of the University of Western Macedonia, under the supervision of Prof. Fragulis Georgios, is exclusively the product of my personal work and does not infringe any intellectual property rights of third parties and is not a product of partial or complete copying.

The sources used are limited to the bibliographic references only. The points where I have used ideas, text, files and/or sources of other authors are clearly indicated in the text with appropriate citations and the relevant reference is included in the bibliographic references section with full description.

Student Signature: .....

Abstract

This thesis presents a comprehensive review of neuroimaging techniques and deep learning approaches for the classification of Alzheimer's disease and dementia. We examine state-of-the-art methodologies for analyzing brain imaging data, including MRI and PET scans, and evaluate various machine learning and deep learning architectures that have been proposed for automated diagnosis and classification of neurodegenerative diseases.

The work covers preprocessing pipelines for neuroimaging data, including skull stripping and intensity normalization, and explores how modern AI techniques such as convolutional neural networks (CNNs) and attention mechanisms can be applied to detect subtle patterns associated with cognitive decline. We also review publicly available datasets commonly used in this domain and discuss the challenges and future directions in AI-assisted dementia diagnosis.

Keywords: Alzheimer's Disease, Dementia, Deep Learning, MRI, PET, Neuroimaging, Classification, Convolutional Neural Networks, Medical Image Analysis

Περίληψη

Η παρούσα διπλωματική εργασία παρουσιάζει μια ολοκληρωμένη επισκόπηση των τεχνικών νευροαπεικόνισης και των προσεγγίσεων βαθιάς μάθησης για την ταξινόμηση της νόσου Αλζχειμερ και της άνοιας.

Λέξεις Κλειδιά: Νόσος Αλζχειμερ, Άνοια, Βαθιά Μάθηση, MPI, ΠΕΤ, Νευροαπεικόνιση, Ταξινόμηση

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Fragulis Georgios, for his guidance and support throughout this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Contextual Background

The evolution of neuroimaging has transformed the diagnostic landscape for Alzheimer's disease. Multiple imaging techniques, including MRI, PET, and CT, are used to support clinical evaluation.

Structural Magnetic Resonance Imaging (MRI) excels at quantifying brain atrophy, particularly in the hippocampus, where volumetric reductions serve as predictors of progression from Mild Cognitive Impairment (MCI) to AD.

Positron Emission Tomography (PET), with amyloid- and tau-specific tracers, enables visualization of amyloid-β (Aβ) plaques and tau pathology (for more on AD pathology, see Appendix 1). Multimodal approaches integrating multiple imaging modalities provide superior sensitivity and specificity for early diagnosis and longitudinal disease tracking. AI- and deep-learning–enhanced multimodal frameworks now fuse neuroimaging data with genetic, clinical, and neuropsychological variables, achieving diagnostic accuracies exceeding 95% in diverse cohorts [7].

Neuroimaging allows for a non-invasive way to produce quantifiable and objective measurements. The combination of multiple modalities can help to detect structural changes, hypometabolism , connectivity loss and even tau or amyloid deposition. At a time where new treatments emerge at the preclincal stage the use of neuroimaging can provide objective information about the progression stages and the efficacy of the treatment. [8]

The latest innovations in imaging technologies and the impact on the diagnosis of AD have shifted diagnoistic criterial. According to the Naitonal Insitute of Aging (NIA) and the Alzheimer's Association staging is based on amyloid , teu and neurodegeneration

profiles. Core-1 biomarkers include low CSF Aβ42/40 ratios or positive amyloid PET, with Core-2 Biomarkers tracking progression - elevated tau PET or plasma p-tau217) . MRI protocols like diffusion tensor imaging and quantitative susceptibility mapping are used for enthorhinal-hippocampla circuits where early tau deposition can begin even up to two decades early. [8]

For an expanded overview of AD pathology, see Appendix 1.

AD develops slilently for 15-20 years, making early identification and diagnosis crucial. The biomarkers of tau PET and plasma p-tau217 trace early tau sprad and predict neurodegeneration. Cognitive dysfunction for patients with amnesic MCI progresses to AD at a rate of 12-15% annually compared to 1-2% for cognitively normal adults . Thus imaging-based tracking of amyloid and tau staging (Braak Stages) helps identify individuals at highest risk. [9]

1.2 Research Significance

Dementia poses a huge challenge for aged individuals , even more so with a very high percentage of aging population [**?**]. In the US , about 7.2 million individuals are living with AD as of 2025 , a number expected to double by 2060 (13.8 million) . The disease additionally creates a massive economic burden on healthcare systems and caregiving costs , reaching 384 billion dollars , far exceeding other chronic conditions. These trends highlight the need for improved diagnostic capabilities and treatment strategies. [10]

Understanding lifetime dementia risk is essential for prevention and resource planning. One study estimated a 42% baseline risk from age 55 to 95 for developing dementia, with higher prevalence observed in women (45–60%), Black adults, and carriers of the *APOE* ε4 allele [11].

Another interesting aspect that can help prevention and enhance resource planning estimates is the understandning of lifetime dementia. In one study a 42% baseline risk from age 55-95 was found , with higher prevalence in women (45–60%), Black adults, and carriers of the *APOE* ε4 allele [11].

Rougly 697 cases of dementia exist per 10.00 people over 50 as reported in [12] . From those 324 cases are of AD and 115 from vascular dementia. Also in this study it is mentioned that women present greater rates and that prevalence doubles every two years. As is understood population aging continues to be the most important cause for increased rates of dementia , despite trends not deviating significantly from historical statistics.

In order to enhance patient care, clinical decision-making, and research, highly precise diagnostic techniques are necessary for dementia in general and Alzheimer's disease (AD) in particular. In order to diagnose AD-related brain disease earlier and develop more individualized treatment plans, neuroimaging is a key technique [**?**].

1.3 Objective of the Review

The goal of this review is to summarize existing work and provide an intuitive explanation of each step in the pipeline from the imaging tools to the algorithms used for classification, discuss existing image datasets available and highlight gaps for future research.

Chapter 2

Overview of Dementia and Alzheimer's disease imaging

2.1 Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging (MRI) is a non-invasive method for visualizing the internal structure and certain functional aspects of the human body. In contrast with other imaging modalities such as CT scans, MRI uses a large magnetic field and non-ionizing electromagnetic radiation, and therefore does not cause exposure-related hazards. Detailed imaging occurs through the use of radio-frequency (RF) pulses in the presence of a carefully controlled magnetic field, producing high-quality images. When the patient is placed inside this magnetic field, the nuclei within the body align with the applied field and interact with the electromagnetic pulses. As analyzed below, the atoms precess, and this precessional movement is detected by receiver coils.

2.1.1 Spin

Atomic nuclei possess an intrinsic property known as *spin*. Spin occurs in multiples of 1/2 and may be positive or negative. Not all particles possess spin; it appears in particles with an odd number of protons and neutrons ($P + N = \text{odd}$) [13].

2.1.2 Properties of Spin

When a particle with net spin is placed within a magnetic field of strength B, it can absorb a photon of frequency ν. This frequency depends on the gyromagnetic ratio γ of the particle:

$$\nu = \gamma B.$$

For hydrogen atoms, the gyromagnetic ratio is

$$\gamma = 42.58 \text{ MHz/T}.$$

MRI uses nuclei with either an unpaired proton or an unpaired neutron, allowing them to possess a net spinning charge—i.e., angular momentum. Because spin is associated with electric charge, these nuclei generate a magnetic field and behave as magnetic dipoles.

2.1.3 Hydrogen Nuclei in MRI

Hydrogen nuclei are the most abundant MR-visible nuclei in the human body (primarily in water, H_2O) and consist of a single proton with non-zero spin, allowing them to act as magnetic dipoles under a strong magnetic field. When a magnetic field is applied, all hydrogen spin angular momentum vectors tend to align either parallel (spin-up) or anti-parallel (spin-down) to the field. Since spin projection comes in quantized these energy states are distinct. Nuclei can transition between different energy states by absorbing or emitting energy , which in the case of MRI is provided by RF pulses at the (Larmor) resonance frequency.

The axes of spinning protons do not align perfectly parallel to the applied field; rather, they *precess*. The frequency of precession is called the *Larmor frequency*, which is proportional to the magnetic field strength:

$$\omega = \gamma B.$$

After applying the magnetic field, the sum of all nuclei yields a weak net magnetic moment or magnetization vector (MV) parallel to the field. Applying an RF pulse at the Larmor frequency in the presence of a magnetic gradient selects a slice of tissue and excites the nuclei, causing them to absorb energy and rotate into the plane of the RF pulse.

Longer RF pulses produce greater rotation angles. If the pulse has sufficient intensity, it flips the magnetization vector into the transverse (XY) plane, creating a *90° flip angle* in which all protons precess in phase.

At this moment, an RF signal is induced in a receiver coil. This signal depends on the presence and molecular environment of hydrogen atoms. Tightly bound hydrogen, such as in bone, produces weak or no usable signal, whereas loosely bound hydrogen in soft tissues produces a strong signal. The concentration of loosely bound hydrogen is known as the *proton density* or *spin density*.

2.1.4 Relaxation

When the RF pulse is turned off:

- The excited nuclei return to their lower-energy spin state, emitting energy. This decay of transverse magnetization is detected as the *free induction decay* (FID).
- Nuclei realign with the main magnetic field through energy transfer to surrounding molecules. This is the basis of *longitudinal* or *spin-lattice* relaxation.

The time constant describing the recovery of longitudinal magnetization is called the *T1 relaxation time*. Typical values are approximately 500 ms for short T1 tissues. The decay of transverse magnetization, governed by loss of phase coherence among precessing protons, is described by the *T2 relaxation time* (e.g., $\sim$ 80 ms for many tissues).

2.1.5 T1- and T2-Weighted Imaging

Relaxation properties allow distinctions between tissue types. MRI commonly uses two contrast mechanisms:

T1-weighted imaging uses short repetition times(TR) and is detecting T1 relaxation. Tissues that appear bright have short T1 values (e.g. fat) , while those with longer T1 values appear darker . This modality is specifically useful for visualizing anatomical detail due to its characteristic of high spatial resolution.

T2-Weghted Imaging uses long repetition times (TR) and long echo times (TE) . Tissues that are water rich or are abnormal usually have long T2 values and appear bright. [**?**]

By changing TR, TE values and the spatial localization of excitation using gradient fiels, MRI can provide great anatomical detail while maintaining sensitivity to underlying pathological changes. These contrast mechansims are the foundation of MRI and its use in clinical settings.

2.2 The Clinical Use of Structural MRI in Alzheimer Disease

In the case of Alzheimer's disease, we use the features of structural imaging that MRI can produce, and it has become an integral part of the clinical assessment process. New data from clinical studies showing changes in structural markers from preclinical to overt stages of the disease are reshaping the diagnostic landscape and influencing future diagnosis and treatment.

In AD Alzhimer can help find structural changes that take place in the brain of a patient , and it is has become routine , with MRI being one of the most commonly used modalities among clinical settings . As has been mentioned before structural brain changes can take place even up to 20 years before without cognitive decline symptoms and become more obvious and accelerate decline as time goes on. One of the earliest and important signs is shrinkage in the medial temporal areas of the brain, which is used to now diagnose MCI (Mild Cognitive Impairment) . MRI can also allow to distinguish between AD and other types of dementia. Aditionally , shrinkage in the hippocampus , or in the brain matter as a whole can be a sign that neurodegeneration is already underway , marking an important step in the progress of the disease.

MRI is also increasingly used in research to track the effectiveness of disease-modifying drugs based on the structural markers mentioned above. While MRI plays an important role in tracking disease progression and determining atrophy in specific brain regions, it is essential to point out that many other imaging and non-imaging techniques are used in clinical assessment.

The utility of structural imaging and other markers will be enhanced by standardization of acquisition and analysis methods, as well as by the development of robust algorithms for automated assessment [14].

As previously mentioned, Alzheimer's disease is associated with progressive accumulation of abnormal proteins in the brain, which lead to synaptic, neuronal, and axonal damage. Neurobiological changes occur years before symptoms appear, following a stereotypical pattern: early medial temporal lobe involvement (entorhinal cortex and hippocampus), followed by progressive neocortical damage [**? ?**].

The delay in cognitive impact suggests that the toxic effects of abnormal proteins accumulate progressively until they reach a critical threshold where cognitive symptoms become noticeable. For example, amnesic MCI is followed by other cognitive deficits across multiple domains until a disability threshold is reached and typical diagnostic criteria for AD are fulfilled.

Due to new research on disease-modifying drugs that aim to slow disease progression, the value of identifying individuals at earlier stages has increased substantially. Several studies have shown correlations between tissue damage or loss in characteristically vulnerable regions—such as the hippocampus and entorhinal cortex—and the likelihood of progression from MCI to AD. This is also the main objective of the current work: to determine whether algorithms can reliably predict this conversion, which would have a major impact on treatment and disease management [14].

2.3 Atrophy as a Neurodegeneration Marker

Atrophy is a consequence of progressive neurodegeneration. Structural changes can be mapped onto the stages of tangle deposition based on the Braak staging system, as well as onto specific neuropsychological deficits. The earliest signs occur in the perforant pathway, producing memory impairments. Later changes in the parietal and frontal neocortices correspond to language deficits, visuospatial impairments, and behavioral changes.

Changes in structural measures such as whole-brain volume, entorhinal and hippocampal volume, temporal lobe volume, and ventricular enlargement can be identified using MRI, and these measures may serve as potential biomarkers. To be used clinically, such biomarkers must be identifiable across all patients and provide clear distinctions between disease stages.

In the asymptomatic stage, amyloid markers provide indirect evidence of disease pathology, whereas after the onset of Mild Cognitive Impairment (MCI), structural changes are more sensitive indicators [14].

2.4 Alzheimer Disease Criteria and MRI

The report from the International Working Group (IWG-2) Criteria for Alzheimer's Disease defines a shift from viewing AD as a clinicopathological entity to a clinicobiological one. The new diagnostic framework requires both an appropriate clinical phenotype (typical or atypical) and the presence of a biomarker consistent with AD pathology.

The criteria cover the full range of disease stages, including typical AD, atypical variants, mixed AD, and preclinical states (asymptomatic at-risk and presymptomatic AD).

Pathophysiological markers are restricted to those that directly indicate amyloid or tau pathology and must be present for a diagnosis of AD.

Specific Criteria

- A specific CSF signature (decreased $A\beta_{42}$ together with increased T-tau or P-tau concentrations).
- Increased tracer retention on amyloid PET.
- Presence of an autosomal dominant AD mutation.

Topographical biomarkers such as volumetric MRI (e.g., hippocampal atrophy) and FDG-PET have been removed from the core diagnostic algorithm, as they lack sufficient pathological specificity for AD detection. Their new role is to monitor disease progression over time [**?**].

Despite this change, hippocampal atrophy remains one of the most established and validated MRI markers of AD. In vivo measurements correlate with Braak staging and neuronal counts, with volume reductions varying by disease stage (e.g., 15–30% in mild dementia, 10–15% in MCI with mild dementia) [14].

2.5 Computed Tomography (CT)

CT scans are a useful tool in the diagnosis of Alzheimer's disease (AD), providing images of anatomical structures in brain regions such as the medial temporal lobe, where atrophy is considered a marker for AD conversion. However, during the diagnostic process of cognitive complaints or deficiencies, MRI scans are generally preferred due to their superior soft-tissue contrast. CT scans are used primarily when MRI presents contraindications (e.g., pacemakers) [15].

The use of CT in combination with nuclear imaging techniques has increased, first with the introduction of PET/CT and later SPECT/CT in AD diagnosis. These combined modalities emerged because the anatomical localization of functional abnormalities in nuclear imaging alone was often imprecise. Adding CT provides accurate anatomical reference and resolves localization issues, provided the modalities are properly coregistered [**?**].

Given this context, the review of the basic principles of CT will remain brief, since this work does not focus on CT images directly, nor do CT scans provide structural detail comparable to MRI. Nevertheless, CT remains an important imaging method and should be acknowledged.

2.5.1 Mechanics of a CT Scan

2.5.2 Historical Development of CT

The discovery of X-ray radiation is attributed to Wilhelm Conrad Röntgen, who performed the first cathode-tube experiments on November 8, 1895. X-rays possess special properties: they can energize atoms (producing photons via fluorescence), and they can penetrate opaque materials.

Initially, X-ray projection imaging captured a two-dimensional image of a three-dimensional object. However, important structural information was lost due to overlapping tissues, while low-contrast regions were difficult to distinguish. Additionally, scattering produced noise and degraded image quality.

The term *computed tomography* reflects its two essential components: "computed," referring to numerical reconstruction, and "tomography," meaning cross-sectional slicing. Modern CT scanners use X-ray energies between 100–150 kV.

Image reconstruction in CT relies on the mathematical foundations of the Radon transform and its inverse. Radon demonstrated that a 2D image can be reconstructed from a set of projections collected at multiple rotation angles.

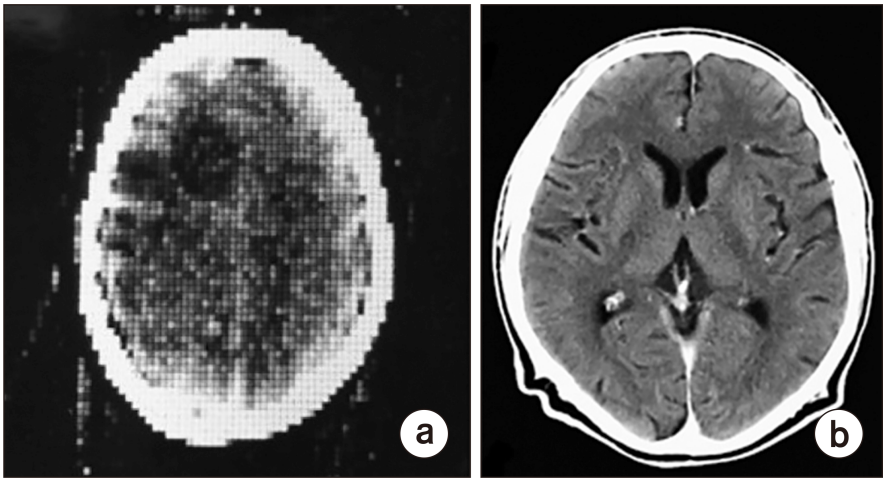

FIGURE 2.1: (a) The first CT image produced at Atkinson Hospital. (b) A modern CT scan from an advanced scanner [1].

2.5.3 Basic Physical Principles of CT

X-ray Attenuation

After X-rays pass through an object, some photons are absorbed while others reach the detector. X-ray attenuation follows the exponential law:

$$I_x = I_0\, e^{-\mu x}, \tag{2.1}$$

where I_0 is the initial intensity, I_x the transmitted intensity, x the material thickness, and μ the linear attenuation coefficient.

For multiple materials along the path:

$$I = I_0 \exp\left[-(\mu_1 x_1 + \mu_2 x_2 + \cdots)\right], \tag{2.2}$$

known as the Lambert–Beer law.

Attenuation can also be expressed as a line integral:

$$\ln\left(\frac{I}{I_0}\right) = -\int \mu(s)\, ds. \tag{2.3}$$

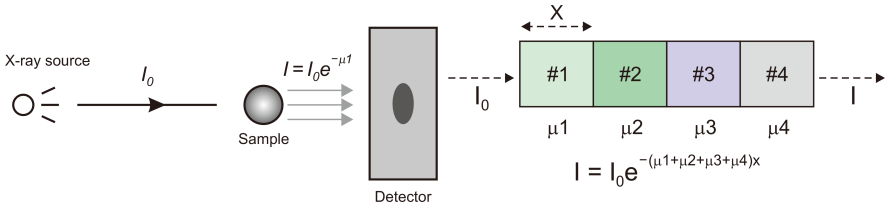

FIGURE 2.2: Illustration of X-ray attenuation through a sample composed of multiple materials. The final intensity corresponds to the cumulative effect of all attenuation coefficients along the beam path.

2.5.4 Data Acquisition

During a CT scan, the sample (e.g., a patient's head) lies on a table while an X-ray source rotates around it. Opposite the source, detectors measure the transmitted X-ray intensity. These measurements are sent to the Data Acquisition System (DAS), which digitizes them and prepares them for reconstruction.

Acquisition criteria:

- Projections must be collected over many angles (typically 360° or 180° with symmetry).
- Each projection must fully include the object.
- The object must remain still.

2.5.5 Image Reconstruction

The goal of CT reconstruction is to compute the 2D attenuation map from the measured 1D projections (line integrals).

For an object described by a function $f(x, y)$, the Radon transform is:

$$p(s, \varphi) = \mathcal{R}f(x, y). \tag{2.4}$$

To convert (x, y) to rotated coordinates (s, u):

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} s \\ u \end{pmatrix}. \tag{2.5}$$

Thus, the projection data are given by:

$$p(s, \varphi) = \int f(s\cos\varphi - u\sin\varphi,\ s\sin\varphi + u\cos\varphi)\, du. \tag{2.6}$$

Early CT systems applied a high-pass filter to sharpen images. Today, reconstruction methods include:

- matrix inversion,
- iterative algorithms,
- Fourier techniques,
- filtered backprojection (FBP),
- 3D Radon-based approaches.

The most common modern method is based on the *central slice theorem*, which links the Radon transform to the 2D Fourier transform.

2.5.6 CT Numbers / Hounsfield Units

A CT image consists of voxels (3D volume elements). Each pixel's intensity reflects the mean attenuation coefficient in its voxel. CT uses a 12-bit grayscale (4096 levels). The Hounsfield Unit (HU) scale normalizes attenuation values relative to water:

$$\text{HU} = 1000 \cdot \frac{\mu_{\text{pixel}} - \mu_{\text{water}}}{\mu_{\text{water}}}. \tag{2.7}$$

2.6 PET / PET-CT

Positron Emission Tomography (PET) is an analytical process in which compounds labeled with radioisotopes are used as molecular probes to image and measure bio-

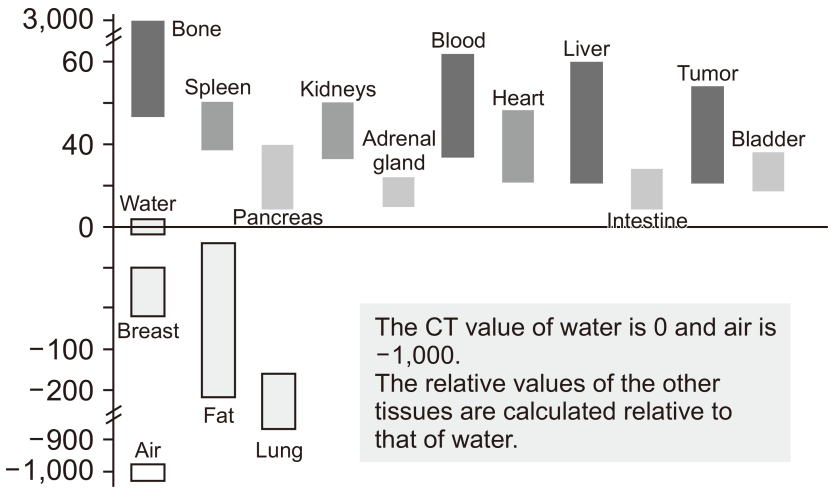

FIGURE 2.3: Hounsfield scale and representative values for different tissues [1].

chemical processes *in vivo*.

2.6.1 Basic Physics

Positron emission is based on proton-rich nuclei, called "emitters," which are unstable. To stabilize themselves, they may either release excess protons and gain neutrons, or capture electrons. The first process is known as *positron emission*, and the second as *electron capture (EC)*. Both processes are isobaric decays, meaning the mass number remains unchanged between the parent and daughter nuclei. In nuclei with low atomic weight, positron emission is more prevalent, while in heavier nuclei electron capture dominates.

The process can be represented as:

$$ {}_{m}X^{n} \longrightarrow {}_{m-1}Y^{n} + \beta^{+} + \nu $$

$$ p \rightarrow n + e^{+} + \nu_{c} $$

Because of the excess protons, a nuclear transmutation occurs, producing a daughter nucleus with the same mass number but an atomic number reduced by one. The emitted neutrino escapes without interacting with the surrounding material.

The emitted positron is highly interactive due to its small mass and positive charge. It travels a short distance and is slowed by scattering in the surrounding tissue.

The distance traveled (the *range*) depends on its energy:

$$E_{\text{positron}} + E_{\text{neutrino}} = \text{transition energy} - 1022 \text{ MeV}$$

As the positron slows, it eventually annihilates with an electron, producing either:

- a pair of 511 keV photons (direct annihilation), or
- a short-lived positronium, which also decays into two 511 keV photons.

These photons, called *annihilation photons*, have energies corresponding to the rest mass of an electron and positron and are emitted nearly 180° apart. However, due to residual momentum, the emission angle varies slightly around 180°.

2.6.2 Detection of Annihilation Radiation

To detect annihilation photons, the system uses a narrow coincidence timing window (3–15 ns). Both photons must be detected within this window for the event to be accepted. This method is called *coincidence detection* or *electronic collimation*.

The interaction is assumed to have occurred somewhere along the straight line connecting the two detectors. This line is known as the *line of response (LOR)* or *coincidence line*.

2.6.3 Factors Limiting the Spatial Resolution of PET

1. **Positron Range:** Because the positron travels a finite distance before annihilation, the detected photons do not originate exactly where the positron was emitted. This introduces an intrinsic spatial uncertainty.
2. **Non-colinearity of Annihilation Photons:** The photons are not emitted at exactly 180°. Small deviations (typically $< 0.5°$) cause blurring. The effect depends on detector ring diameter—about 1 mm for a 50 cm ring and 2 mm for a 90 cm whole-body system.
3. **Detector Size:** The intrinsic spatial resolution depends on the size of the individual detector crystals used in modern PET scanners, which typically consist of small scintillator arrays coupled to larger photodetectors [16].

2.6.4 PET-CT

PET and Anatomic Imaging

A major limitation of PET imaging is its low spatial resolution and lack of anatomical detail, making it difficult to accurately localize lesions with abnormal radiotracer

uptake. Distinguishing physiological from pathological uptake can be challenging.

Initially, non-integrated scanners were used: PET and CT images were acquired separately and then aligned visually or using fusion algorithms. However, differences in patient position and organ motion significantly reduced accuracy. The dual-scan approach is also costly, time-consuming, and uncomfortable for patients.

Integrated PET/CT

PET scanner was introduced in 1998. It has the ability to show both body structure through the CT scan but also body function by utilizing the PET scan. This not only helps doctors make more accurate decisions but also helps patients to make the process faster , easier and cheaper for them. [16**?**].

2.6.5 The Use of PET in Alzheimer's Disease

Alzheimer's disease (AD) is characterized by amyloid plaques, neurofibrillary tangles (intracellular hyperphosphorylated tau aggregates), activated microglia, neurotransmitter changes, and neuronal loss. Brain changes can take place many years before any symptoms appear so the new criteria that have been proposed , have emphasized CSF analysis and brain imagin as early biomarkers and useful tools to track progression. [**?** **?**].

New PET imaging technologies allow the identification of AD in prodromal stages and support the development and monitoring of disease-modifying treatments. PET enables measurement of multiple functional processes in the brain, including metabolism and neurotransmitter activity, providing region-specific insights that strengthen diagnostic accuracy and treatment evaluation.

2.6.6 Processes Assessed by PET

Brain Glucose Metabolism

To measure glucose metabolism, PET uses the tracer 2-fluoro-2-deoxy-glucose (F-FDG). A decline in glucose metabolism occurs long before clinical symptoms appear. This decline is region-specific, most prominently affecting the parietotemporal, frontal, and posterior cortices [**?**].

Patients with AD can be diagnosed with up to 90% sensitivity using FDG-PET [**?**]. Differentiation from other dementias is more difficult.

Longitudinal PET studies show:

- Conversion from healthy to MCI is best predicted by **medial temporal** glucose hypometabolism.
- Conversion from MCI to AD is best predicted by hypometabolism in the **posterior cingulate cortex**.

Functional MRI (fMRI) also helps distinguish patterns of brain activity. Its basis lies in the magnetic properties of deoxyhemoglobin and the fact that blood flow increases more than oxygen metabolism during neural activation. This produces a subtle MR signal increase known as the *blood oxygenation level-dependent (BOLD)* effect. Modern fMRI achieves spatial resolution near 1 mm and temporal resolution around 1 s [17].

Chapter 3

Key Image Datasets for Dementia and Alzheimer's Disease

3.1 ADNI

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is one of the most extensive and significant research programs in the field of Alzheimer's disease. ADNI represents a rare public–private partnership, supported by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the NIH, major pharmaceutical companies (such as Pfizer, Eli Lilly, Merck, GlaxoSmithKline, and AstraZeneca), and nonprofit organizations including the Alzheimer's Association.

The purpose of ADNI, which spans multiple clinical sites across the United States and Canada, is to identify and validate biomarker-based and neuroimaging indicators linked to cognitive and functional decline in aging populations, including individuals with Alzheimer's disease (AD), mild cognitive impairment (MCI), and cognitively healthy controls. ADNI integrates genetic, cerebrospinal fluid (CSF), clinical, and cognitive data with a wide range of imaging modalities, including structural MRI, FDG-PET, amyloid PET, tau PET, and DTI.

The multiple modalities of the dataset in order to provide researchers with data that show brain structure changes and cognitive abilities. It is used to accelearate work on early diagnosis , follow disease progression, and assess treatment efficacy.

The goal of ADNI is to provide standardized datasets with MRI and PET imagin from large cohorts across mulitple centers , and identify sensitive markers , as well as assessing disease efficiency .

This multimodal dataset allows researchers to monitor changes in brain anatomy,

metabolism, and cognitive performance over time for early diagnosis, disease monitoring, and treatment evaluation.

The main objectives of ADNI include the standardization of imaging procedures for long-term, multicenter PET and MRI data collection, the acquisition and validation of biomarkers and clinical metrics in a large participant cohort, and the identification of the most accurate and sensitive markers for diagnosing AD and MCI and evaluating treatment effectiveness. Another major goal is the establishment of a publicly accessible data repository containing comprehensive longitudinal information on clinical outcomes, biomarkers, imaging, and cognition. With its ability to power thousands of studies and substantially advance our understanding of neurodegenerative disease progression, ADNI has become a cornerstone resource for Alzheimer's research worldwide [**?**].

3.2 AIBL

The Australian Imaging, Biomarkers and Lifestyle (AIBL) study is another major initiative designed to collect data from approximately 1000 individuals aged over 60 to enhance ongoing research on Alzheimer's disease. In this study, volunteers participated in a screening interview, underwent cognitive assessment, provided 80 ml of blood, and completed detailed health and lifestyle questionnaires. One quarter of the participants also underwent amyloid PET imaging and MRI scans, while a smaller subgroup of about 10% received ActiGraph activity monitoring and body composition scanning.

In total, the study collected data from 211 patients with clinically confirmed Alzheimer's disease, 133 participants with mild cognitive impairment, and 768 cognitively healthy controls. This final cohort represents a highly motivated and well-characterized group that contributes significantly to ongoing Alzheimer's research. Participants are reassessed at 18-month intervals to determine the predictive utility of biomarkers, cognitive measures, and lifestyle factors as indicators of AD and future cognitive decline. This longitudinal dataset is invaluable for evaluating disease progression, informing predictive models, and establishing clearer labels for different disease stages. Moreover, it highlights the importance of large-scale studies for defining individual variability in both pathological and normal cognitive aging. It also provides critical insights into the MCI stage, which is widely considered one of the most important phases for early detection and targeted intervention [**?**].

3.3 OASIS

The Open Access Series of Imaging Studies (OASIS) is a longitudinal neuroimaging initiative similar in purpose to AIBL. It consists of MRI and PET imaging along with relevant clinical data for 1098 participants collected over a 15-year period. Participants range in age from 42 to 95 years, including 605 cognitively healthy adults and 493 individuals across different stages of cognitive decline.

The OASIS-3 dataset includes over 2000 MRI sessions featuring multiple structural and functional imaging sequences, as well as approximately 1500 raw PET scans and their corresponding post-processed outputs generated through the PET Unified Pipeline (PUP). The dataset also contains additional processed neuroimaging products such as volumetric MRI segmentations and PET-derived metrics. Furthermore, OASIS provides dementia status, APOE genotype information, and longitudinal clinical and cognitive assessments, making it an invaluable resource for research on aging and neurodegenerative disease [**?**].

Chapter 4

Pre-Processing and Feature Extraction Techniques

4.1 Intensity Normalization

4.1.1 Introduction

The accuracy of automated processing algorithms for Magnetic Resonance (MR) images is of immense importance for clinical diagnosis and disease evaluation in progressive diseases like Alzheimer's. However, the reliability of processing applications—such as registration, segmentation, and comparison—requires a standardized intensity value scale that accurately depicts the underlying tissue type. Variability is inherent in MR images, as scanner-related artifacts, noise, differing acquisition protocols, and other parameters can significantly affect the obtained intensity values. Additionally, challenges such as low contrast and partial volume effects degrade the performance of these algorithms [2].

To mitigate these issues, intensity normalization is employed as a critical pre-processing step. The goal is to create a standardized scale that accurately reflects the underlying tissue type by unifying the mean and variance within an MR image. The efficacy of this step directly influences the performance of longitudinal processes essential for tracking disease progression [2, 18, 19].

Its primary objective is to transform the intensities of a set of images to a standardized scale. Ideally, scanner-related differences should not depict the same tissue type with different intensity values; conversely, the same underlying biological value should yield consistent intensity values across all setups [20, 21]. This process addresses the inherent intensity variability in MR acquisition and is essential for the reliability and accuracy

of downstream algorithms [22, 23].

4.1.2 Necessity and Significance

The lack of tissue-specific intensity standardization for MR images creates variability, resulting in differences between two scans of the same subject, even when performed on the same scanner with the same pulse sequence [2]. As noted previously, this reduces the accuracy of subsequent processing algorithms. Key factors identified in the literature include:

- **Scanner Related Artifacts and Noise:** Inherent scanner artifacts and noise must be addressed as they interfere with automated processing.
- **Low Contrast and Partial Volume Effects:** Ideally, edges between brain tissues (e.g., Cerebrospinal Fluid (CSF), White Matter (WM), Gray Matter (GM)) should be distinct. However, low contrast and partial volume effects—where multiple tissue types within a single voxel are averaged—blur these boundaries [**?**].
- **Intensity Variations:** Different acquisition protocols, scanner differences, and heterogeneity of data sources contribute to intensity variations [24]. These variations occur both between different subjects and within sequential scans of the same patient [25].
- **Lack of Specific Tissue Meaning:** Unlike Computed Tomography (CT), which uses standardized Hounsfield Units, MRI intensity values are arbitrary and dependent on the specific scanner setup. This severely affects longitudinal and quantitative analysis [23].

4.1.3 Goals and Principles

The normalization protocol adheres to the SPIN principles introduced by Shinohara et al. [21]. Unlike methods that transform images based on a learned standardized histogram [22], this approach proposes transforming images based on the white matter variance unique to every scan. The primary goals are:

1. **Standardized Interpretation:** To create an accurate representation of tissue types across multiple setups and acquisition protocols.
2. **Reduced Variability:** To minimize the discrepancy between intensity distributions across subjects and visits within tissue classes.

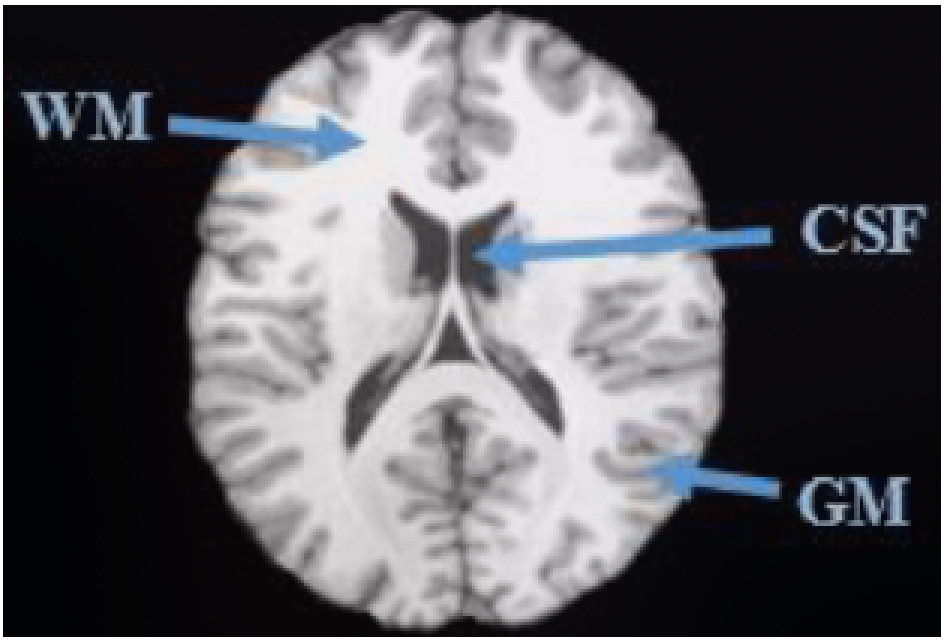

FIGURE 4.1: Brain image depicting artifacts present in MR images and different tissue types. Adapted from [2].

3. **Biological Interpretability:** To yield biologically interpretable units, facilitating quantitative studies.

4. **Robustness to Pathology and Artifacts:** To perform reliably even in the presence of pathology (e.g., lesions) or imaging artifacts.

4.1.4 Categories of Intensity Normalization Methods

Methods can be broadly categorized into:

1. Classical/Statistical Methods

2. Feature-Based Harmonization

3. Deep Learning-Based Approaches

Classical Methods

1. Min-Max Normalization This technique utilizes the full range of intensity values to reduce variance and implement a standard range, typically between 0 and 1 [26, 27].

$$v' = \frac{v - \min(A)}{\max(A) - \min(A)} \tag{4.1}$$

Where v' is the normalized value, v is the original data point, and $\min(A)$ and $\max(A)$ are the minimum and maximum feature values, respectively.

2. Z-Score Normalization This method transforms each intensity value by subtracting the mean and dividing by the standard deviation (excluding background). It represents how many standard deviations a score is from the mean [24]. A disadvantage is that high intensities (e.g., contrast agents or artifacts) may be attenuated, risking information loss.

$$\mu = \frac{1}{|B|} \sum_{b \in B} I(b) \tag{4.2}$$

$$\sigma = \sqrt{\frac{\sum_{b \in B} (I(b) - \mu)^2}{|B| - 1}} \tag{4.3}$$

$$I_{\text{norm}}(x) = \frac{I(x) - \mu}{\sigma} \tag{4.4}$$

3. Mean and Standard Deviation Setting ($\mu \pm 3\sigma$ / M-std) This approach sets the mean and standard deviation to fixed values (e.g., $\mu = 0, \sigma = 1$) [28]. Alternatively, it limits intensity dynamics to $\mu \pm 3\sigma$ [29]. This is effective when the underlying distribution is Gaussian; however, non-Gaussian distributions may suffer from gray value truncation [25].

4. Percentile Method A robust method using specific percentiles (e.g., 5th and 95th) as minimum and maximum values to remove outliers [28].

5. Histogram Matching Widely used, this transforms the image histogram to match a reference histogram [22].

- **Nyul and Udupa's Method:** Involves a *Training Step*, where landmarks (percentiles, modes) are learned from a training set to form a standard histogram, and a *Transformation Step*, where new images are mapped to this scale using piecewise linear transformation.

- **Joint Histogram Matching:** Uses information-centric criteria to analyze pixel pair relationships between input and reference images. While potentially more precise, it is computationally intensive [18].

6. Homomorphic Filtering This technique leverages the illumination-reflectance model, where intensity $I(x, y)$ is the product of illumination L and reflectance R:

$$I(x, y) = L(x, y) \cdot R(x, y) \tag{4.5}$$

By applying a logarithmic transform, the multiplication becomes addition:

$$\ln(I) = \ln(L) + \ln(R) \tag{4.6}$$

Fourier transform and high-pass filtering are then used to remove low-frequency illumination inconsistencies while preserving high-frequency reflectance components (edges) [2].

7. Gaussian Filtering Used for smoothing, this requires careful selection of the σ parameter. A small σ may be inefficient for normalization, while a large σ can cause loss of edge information.

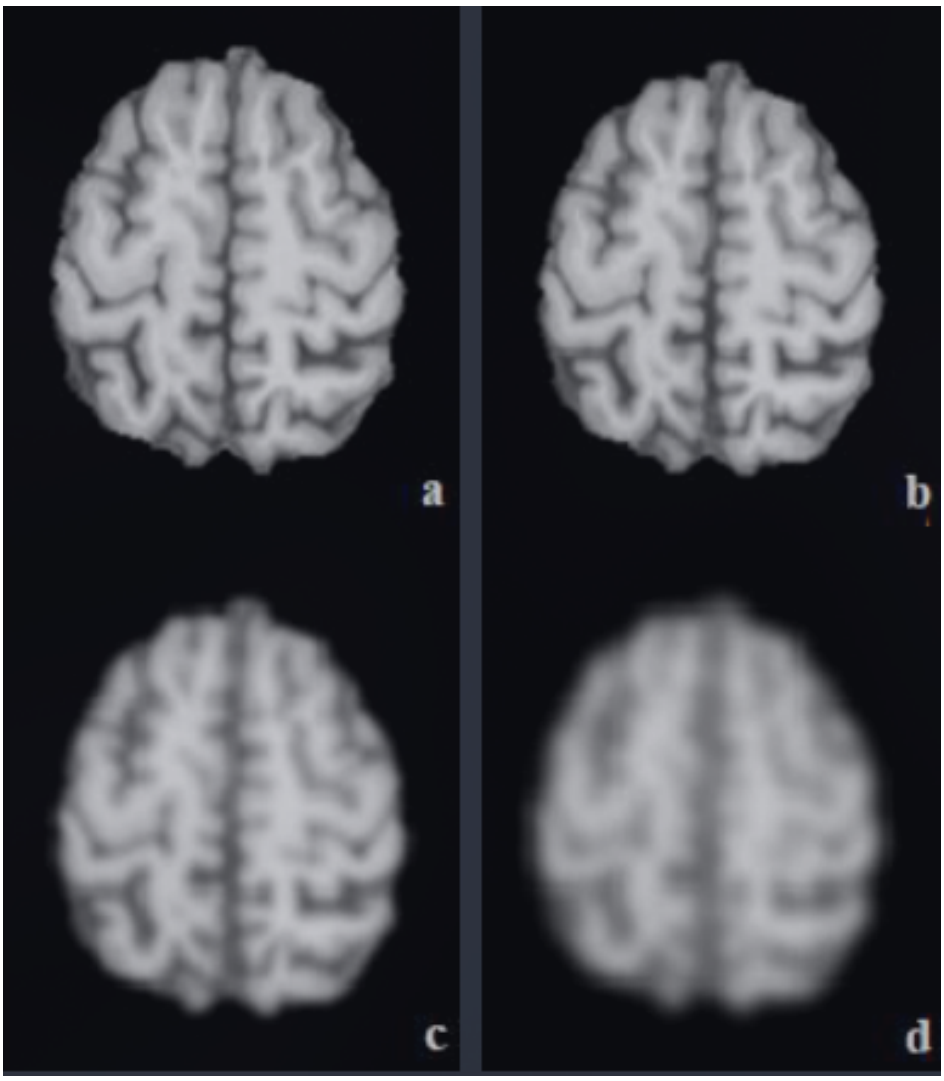

FIGURE 4.2: Original image (a); Image obtained by Gaussian filtering with increasing σ values (b-d).

8. White Stripe Normalization Proposed by Shinohara et al. [21], this method normalizes based on Normal-Appearing White Matter (NAWM). It identifies the NAWM

distribution ("white stripe") and matches its moments (mean and standard deviation) across subjects, providing biologically interpretable units robust to pathology.

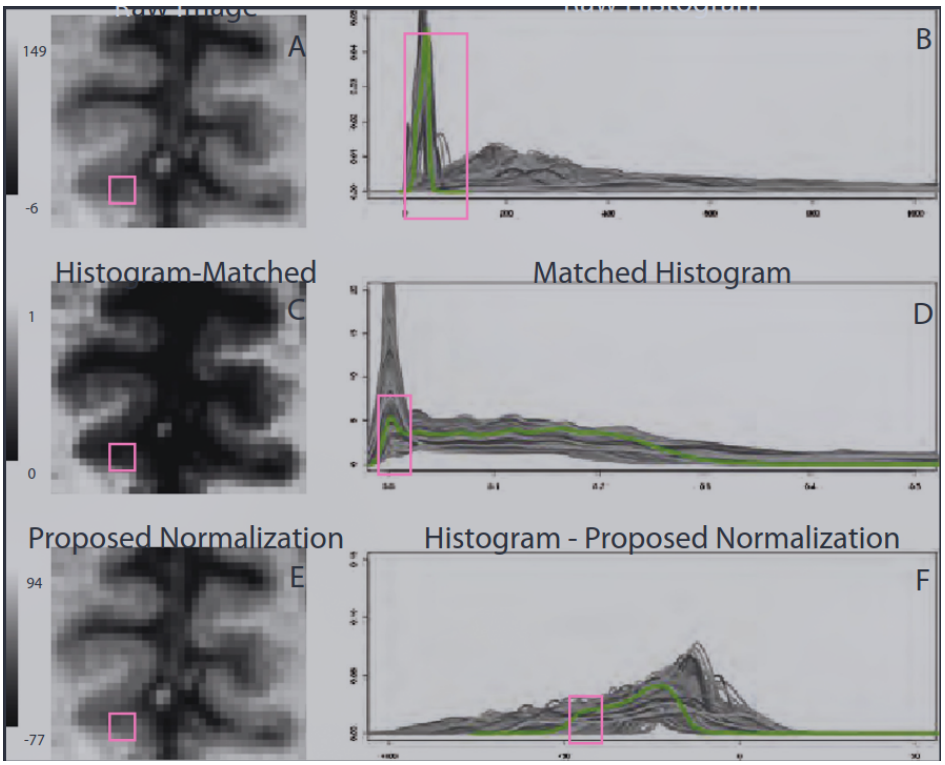

FIGURE 4.3: White Stripe Normalization concept.

9. Fuzzy C-Means (FCM) and Kernel Density Estimation (KDE) Fuzzy clustering assigns data points a membership function (0 to 1) indicating proximity to a cluster center. Normalization uses tissue segmentation to scale intensity relative to tissue means.

Single-Tissue Normalization: Normalize voxel x by the mean intensity of tissue T:

$$I_{\text{norm}}(x) = \frac{I(x)}{\mu_T} \tag{4.7}$$

Two-Tissue (Min-Max) Normalization: Using two tissue means μ_1 and μ_2, define $a = \min(\mu_1, \mu_2)$ and $b = \max(\mu_1, \mu_2)$. The image is rescaled via:

$$I_{\text{norm}}(x) = \frac{I(x) - a}{b - a} \tag{4.8}$$

Kernel Density Estimation (KDE): KDE approximates the probability distribution function to smooth the histogram and locate peaks (e.g., white matter peak π) for

normalization:

$$p(x) = \frac{1}{NMLh} \sum_{i=1}^{NML} K\left(\frac{x - x_i}{h}\right) \tag{4.9}$$

The image is then normalized as $I_{\text{norm}}(x) = c \cdot \frac{I(x)}{\pi}$.

Feature-Based Harmonization

ComBat Originally for gene expression, ComBat is a feature-based method applied to imaging to eliminate batch effects [30]. It assumes feature similarity and allows only for shift or spread in the data.

Deep Learning Methods

Deep learning has emerged as a powerful tool for normalization, particularly in multi-site analysis.

1. Autoencoders (AEs) and Generative Adversarial Networks (GANs) An autoencoder is an algorithm designed to learn an informative representation of data by reconstructing input observations [31]. It encodes input x_i into a latent representation $h_i = g(x_i)$ and reconstructs it via $X_i = f(h_i)$. Training minimizes the reconstruction error:

$$\arg\min_{f,g} \Delta(X_i, f(g(x_i))) \tag{4.10}$$

GANs consist of two networks: a Generator that creates data and a Discriminator that evaluates it. In the context of normalization, authors have proposed using an autoencoder as the generator within a GAN framework. Through adversarial training, the model optimizes to reconstruct images that preserve anatomical features while obscuring scanner-specific biases (detected by the discriminator), improving multi-site generalization [32, 33].

2. Adversarial and Task-Driven Normalization This approach generates images optimized not just for visual similarity, but for specific downstream tasks like segmentation. Delisle et al. [32] utilize two 3D CNNs: a normalization generator and a segmentation network. A discriminator acts as a domain classifier. By integrating the specific task (segmentation) into the loss function, the pipeline can often outperform generic domain adaptation, as the normalization is tailored to maximize segmentation accuracy rather than just histogram matching.

3. Disentanglement Networks Unsupervised image-to-image translation often assumes a one-to-one mapping. To address limitations, frameworks like MUNIT de-

compose image distribution into a *content code* (domain-invariant structure) and a *style code* (domain-specific appearance). By disentangling these, the network can normalize content across arbitrary source domains [33].

4.1.5 Impact of Intensity Normalization

Image Registration Normalization enhances similarity metrics used in registration by ensuring consistent intensity values for similar tissues, leading to improved alignment accuracy [18, 20].

Image Segmentation Intensity variation undermines segmentation and volume measurement. Homogeneous intensities acquired via normalization result in better tissue separation and more robust segmentation, particularly for multi-site data [**?**].

Texture Classification/Radiomics Texture analysis quantifies spatial gray-level variations. Normalization is critical here, as these variations are highly sensitive to acquisition protocols. However, care must be taken, as normalization methods themselves can be sequence-dependent [25].

Brain Template Construction Normalization facilitates the efficient calculation of volume statistics for GM, WM, and CSF. Post-normalization brain templates show clearer tissue separation, making anatomical definitions more distinct [18].

4.1.6 Challenges and Considerations

While critical, intensity normalization faces challenges. The choice of method is often sequence-dependent [24]. Furthermore, deep learning methods, while promising, are limited by the availability of diverse training datasets. As noted by Albert et al. [28], deep learning outperformed classical methods only in specific cases, likely due to data scarcity. The selection of an appropriate strategy remains pivotal for any MR analysis pipeline.

4.2 Denoising

Denoising is an important step in the pre-processing of images with the goal of using them as training data for a learning algorithm. Noise and artifacts can significantly impact the performance of such algorithms, which is why denoising is often included as a standard pre-processing step—especially when the quality of the raw images varies or when data come from multiple scanners (inherent variability).

In denoising, we assume there exists an "ideal" clean image represented as a vector $x \in \mathbb{R}^N$. However, what we observe is a noisy measurement y, which contains noise.

The most common mathematical noise model, due to the Central Limit Theorem and ease of use, is Additive White Gaussian Noise (AWGN). The relationship is:

$$y = x + v.$$

In this model, the noise is assumed to be drawn from a Gaussian distribution with mean zero and variance σ^2:

$$v \sim \mathcal{N}(0, \sigma^2 I).$$

This means that each pixel is corrupted by a random value sampled from that distribution.

The goal is to find a function D such that $\hat{x} = D(y, \sigma)$, producing an estimate as close as possible to the original x.

Since we know the noisy image and the noise distribution, we can formulate the problem using Bayesian inference. We aim to maximize the posterior probability:

$$p(x \mid y) = \frac{p(y \mid x)\, p(x)}{p(y)}.$$

Taking the negative logarithm turns the product into a sum, resulting in the minimization objective:

$$-\log p(x \mid y) \propto \underbrace{-\log p(y \mid x)}_{\text{Data Fidelity}} + \underbrace{-\log p(x)}_{\text{Prior}}.$$

Because of the Gaussian noise model, the likelihood is:

$$p(y \mid x) \propto \exp\left(-\frac{\|y - x\|^2}{2\sigma^2}\right).$$

Taking the negative log removes the exponential:

$$-\log p(y \mid x) = \frac{1}{2\sigma^2}\|y - x\|^2.$$

The second term, $-\log p(x)$, serves as a prior and becomes a regularization term. The

denoising objective becomes:

$$\hat{x} = \arg\min_x \underbrace{\|y - x\|^2}_{\text{Match the Data}} + \lambda \cdot \underbrace{\rho(x)}_{\text{Be a Real Image}} .$$

This formulation motivated the field's "evolution of priors," producing many classical denoising approaches.

Two of the most important classical algorithms are NLM and BM3D.

4.2.1 Non-Local Means (NLM)

NLM grew from the idea of *spatial smoothness*. Earlier algorithms assumed images should be spatially homogeneous, so averaging pixels in local neighborhoods could remove noise while preserving signal. These methods often produced blurry results, especially near edges.

NLM instead uses neighborhoods (patches) of pixels and the concept of *self-similarity* across the entire image. For each pixel p, a patch P surrounding it is compared to patches Q around every other pixel q, using a similarity measure such as Euclidean distance $\|P - Q\|^2$.

The denoised estimate averages all pixels, but with weights proportional to patch similarity:

$$\hat{x}(p) = \frac{\sum_q w(p,q)\, y(q)}{\sum_q w(p,q)}.$$

4.2.2 BM3D

After NLM, the creation of algorithms such as BM3D improved denoising performance so dramatically that researchers questioned whether denoising had reached a theoretical limit.

BM3D stands for *Block-Matching and 3D Filtering*. Like NLM, it finds similar patches, but instead of averaging them directly, it stacks them into a 3D group. A transform (3D wavelet or DCT) is applied so that signal becomes concentrated in a few large coefficients while noise spreads into many small ones. Thresholding the small coefficients and inverting the transform yields high-quality results.

4.2.3 Deep Learning Era

Before deep learning, humans designed the transforms (e.g., DCT) used to enforce sparsity. With deep learning, the network learns the appropriate transform directly

from data.

We define a parametric function f_θ (a neural network) with parameters θ. The network takes a noisy image y and outputs a clean estimate x.

For supervised learning with an MSE loss:

$$\min_\theta \sum_{i=1}^{N} \|f_\theta(y_i) - x_i\|^2.$$

Since clean images are not always available, a common strategy is to corrupt clean images with Gaussian noise and train the model to recover them.

No single architecture is universally best; much progress arises from empirical experimentation to identify the best-performing denoiser.

A recently expanding research direction, highlighted in [34], uses denoising network priors as generative models capable of synthesizing new images. Using iterative methods such as Langevin dynamics, these priors can generate realistic images from noise. While not the primary focus of this work, such techniques can be beneficial in domains like medical imaging, where data scarcity is a major challenge.

4.3 Skull Stripping

High-resolution MRI images contain non-brain tissues such as skin, fat, muscle, neck structures, and eyeballs, unlike other modalities such as PET scans. The presence of these non-brain tissues can severely impact automated processing algorithms such as image segmentation and analysis techniques. Therefore, quantitative morphometric studies of MR brain images—such as those used in Alzheimer's disease research—commonly require a preprocessing step to remove non-brain and extra-cranial tissues [3, 35].

The MRI system produces brain images as 3D volumetric data composed of 2D slices. Further computer-aided processing is essential in order to extract meaningful information, whether for research, diagnostic, or clinical purposes.

As noted earlier, quantitative morphometric studies require isolating brain tissue from non-brain tissue, a process referred to as *skull stripping*. Automated skull-stripping enhances segmentation accuracy, making manual or automated segmentation methods more reliable [4]. Examples from [3] are shown in Figure 4.4.

Voxel-based morphometry (VBM) results also showed significant improvements after skull stripping was applied [4]. Many brain imaging applications benefit from the

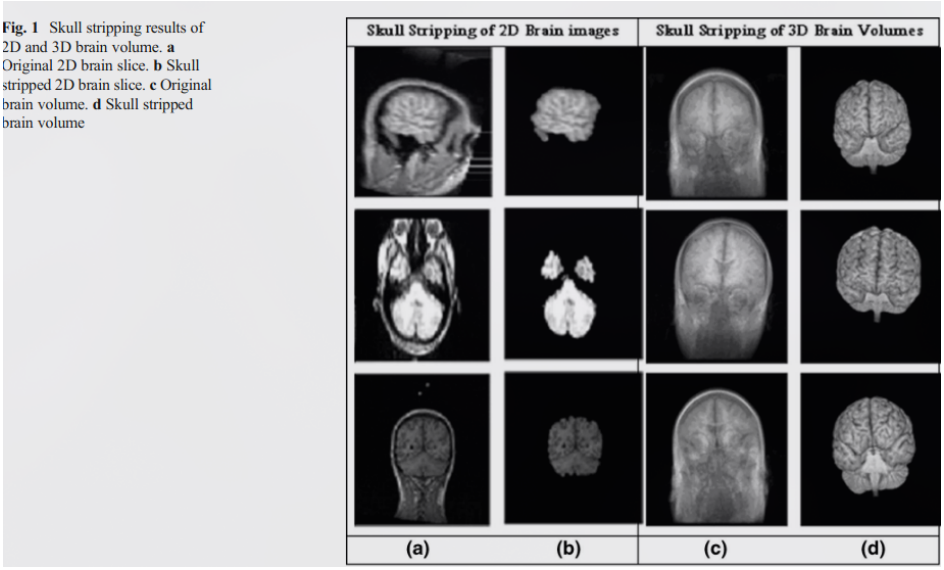

FIGURE 4.4: Examples of automated skull stripping from [3].

precise segmentation of the brain from surrounding tissues, including cortical surface reconstruction, tissue classification, and registration to standard templates [36, 37].

Skull-stripping algorithms are typically evaluated by their speed and their impact on downstream automated tasks such as segmentation [38].

Multiple skull-stripping techniques have been developed due to their success and effectiveness in improving diagnostic and prognostic accuracy [3, 35].

Manual brain segmentation is considered the most robust and accurate approach, but it is extremely laborious, time-consuming, and subject to inter-clinician variability [39]. Manual masks are often treated as the ground truth against which automated methods are validated [4]. This has led to a strong need for automated skull stripping.

Despite significant progress, many skull-stripping approaches still perform well only on specific datasets and often require tuning of hyperparameters [38, 40]. According to [4], skull-stripping methods fall into three primary categories:

1. Manual skull stripping
2. Classical approaches
3. Deep learning–based approaches

4.3.1 Skull Stripping Methods

Based on [3], skull-stripping techniques can be grouped as follows:

- Mathematical morphology-based methods

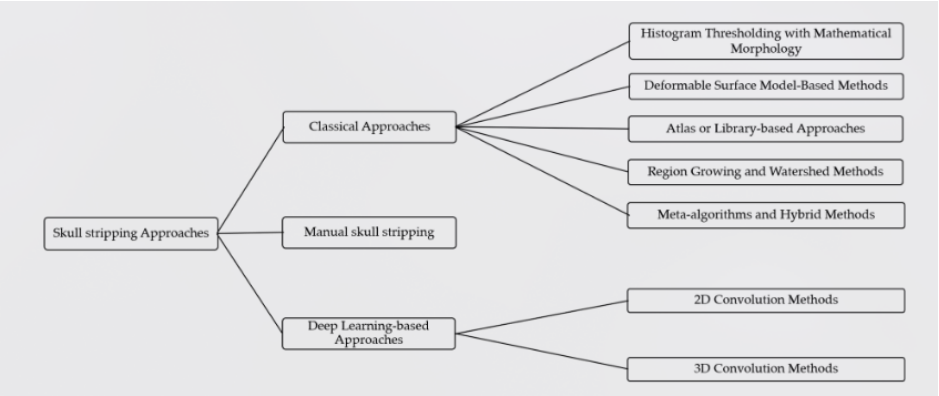

FIGURE 4.5: Overview of skull-stripping categories [4].

- Intensity-based methods
- Deformable surface-based methods
- Atlas-based methods
- Hybrid methods
- Deep learning–based methods

4.3.2 Morphology-Based Methods

These methods use morphological erosion and dilation operations—combined with thresholding and edge detection—to estimate the initial region of interest (ROI) and separate brain from non-brain tissues [3, 41].

A key drawback is that performance depends heavily on empirically tuned parameters related to the shape and size of the morphological structuring elements [35].

Examples

1. **Brain Surface Extraction (BSE).**
   Uses anisotropic diffusion filtering, Marr–Hildreth edge detection, and morphological operations such as erosion and dilation. BSE is fast (≈3.5 seconds) [3, 41].

2. **Brain Extraction Algorithm.**
   Uses diffusion, morphological operations, and connected component analysis for T1W and T2W MRI [4].

3. **SMHASS (2012).**
   Combines deformable models with histogram analysis and morphological preprocessing [3].

4.3.3 Intensity-Based Methods

These methods classify brain vs. non-brain regions using pixel intensities. Examples include histogram-based, region-growing, and edge-based methods. Their primary limitation is sensitivity to MRI noise, low contrast, and bias field artifacts [3, 38].

Examples

1. **Graph Cuts (GCUT, 2010).**
   Uses graph-theoretic segmentation to isolate brain tissue from dura, beginning with thresholding followed by connected submask selection [3].

2. **Region Growing (RG).**
   Forms connected regions based on local intensity similarity. Variants include 2D approaches for coronal slices and MARGA for axial views and low-quality images [3].

4.3.4 Deformable Surface-Based Methods

These methods employ active contours (snakes) that evolve iteratively according to an energy functional. The contour adapts—shrinking or expanding—to match the brain boundary. Level-set formulations provide a robust mathematical framework [3, 36].

Their performance depends strongly on the initial contour and image quality, especially the clarity of edges [42].

Examples

1. **BET (Brain Extraction Tool, 2002).**
   Widely used, fast, and freely available. It expands a tessellated sphere to match brain edges using adaptive forces. BET supports T1-weighted, T2-weighted, and proton density images, processing in 5–20 seconds [42**?** **?**].

2. **Model-Based Level Set (MLS, 2006).**
   Uses curvature and intensity-derived forces to evolve an active contour [3].

3. **3dSkullStrip (AFNI, 2005).**
   BET-like but includes improvements to avoid eyes and ventricles [3, 43].

4.3.5 Atlas / Template-Based Methods

These methods register the subject MRI to one or more anatomical atlases, enabling the transfer of brain masks to the subject [3**?**].

Examples

1. **MAPS (2011).**
   Combines multiple atlas registrations to produce a consensus segmentation [3].

2. **BEaST (2012).**
   Uses nonlocal segmentation with multi-resolution priors. BEaST achieves high accuracy through patch-based label fusion from a library of pre-labelled atlases [39, 44**?**].

3. **Pincram (2015).**
   Employs iterative refinement to propagate labels from multiple atlases [3].

4.3.6 Hybrid Methods

Hybrid approaches combine multiple algorithms or features to improve robustness and accuracy [3, 35].

Examples

1. **SPECTRE (2011).**
   Integrates elastic registration, tissue segmentation, and morphological watershed operations [3].

2. **HWA (Hybrid Watershed Algorithm, 2004).**
   Combines watershed segmentation with deformable surface models. HWA showed the highest sensitivity among compared methods and appeared more robust to parameter changes [36, 44].

3. **BEMA (2004).**
   Runs multiple extractors (BET, BSE, Watershed, etc.) in parallel and merges the results [4].

4. **ROBEX (2011).**
   Combines a Random Forest classifier with a point distribution model to ensure anatomically plausible results. The key advantage is parameter-free operation with consistent performance across datasets—achieving Dice scores of 95.6–97.0% on IBSR, LPBA40, and OASIS without tuning [40, 44].

4.3.7 Deep Learning-Based Methods

Deep learning approaches include 2D and 3D CNNs. While 3D CNNs capture volumetric context, they require higher computational resources. DL methods are generally

categorized as [4, 45]:

- Patch-based CNNs
- Encoder–decoder CNNs (e.g., U-Net)

Encoder–decoder architectures typically perform better, are faster, and can capture global and local features [44].

However despite their performance, deep learning methods showcase several limitations [4, 35]:

- The requirement of large annotated datasets
- The inability of hyperparameter tuning due to the black-box behavior of the networks
- The sensitivity that emerges when trained on healthy individuals

State-of-the-Art Deep Learning Methods

SynthStrip represents a paradigm shift through synthetic training data. The method trains a 3D U-Net entirely on synthetically generated images, randomising intensity distributions, artifacts, and deformations. This yields contrast-agnostic generalisation—a single model processes T1w, T2w, FLAIR, DWI, MRA, CT, and PET images with Dice scores of 96–98%. Processing takes under 2 seconds on GPU [46].

HD-BET (High Definition Brain Extraction Tool) was designed for clinical heterogeneity. Training used 6,586 MRI sequences from 372 patients across 37 European institutions, including brain tumours and resection cavities. The ensemble of five 3D U-Net models with test-time augmentation outperforms BET, 3dSkullStrip, BSE, ROBEX, BEaST, and MONSTR [44].

deepbet achieves the highest reported accuracy for T1-weighted images using a two-stage 3D LinkNet architecture. Trained on 7,837 images from 191 OpenNeuro datasets, deepbet achieves median Dice of 99.0% with processing times of only 0.5 seconds on GPU [47].

Deep MRI Brain Extraction by Kleesiek et al. pioneered CNN-based skull stripping, demonstrating that deep learning could handle pathological brains better than classical methods [45].

4.3.8 Comparative Analysis

Automated vs. Manual Approaches

- Automated methods are faster but may require parameter tuning [38].
- Semi-automated methods are accurate but slow and user-dependent [3].

Evaluation Metrics

Common metrics include [4, 44]:

- Dice coefficient
- Jaccard Index
- Sensitivity / Specificity
- False Positive Rate / False Negative Rate
- Hausdorff Distance
- Average Symmetric Surface Distance (ASSD)

Comparative findings from the literature include [3, 4, 44**?**]:

- McStrip (hybrid) outperforms BET and BSE.
- HWA shows high sensitivity and robustness.
- Deep learning achieves highest Dice and specificity; 3D U-Net has highest sensitivity.
- SynthStrip performs best on pediatric T2-weighted scans, with accuracy increasing with age.
- HD-BET outperformed all classical methods by +1.16 to +2.50 Dice points on the CC-359 multi-vendor dataset.

4.3.9 Multi-Site Dataset Considerations

Large-scale studies aggregate data from numerous imaging sites with different scanners, protocols, and field strengths. Souza et al. created the CC-359 dataset specifically to evaluate vendor and field-strength effects, comparing BET, 3dSkullStrip, FreeSurfer, BSE, ROBEX, BEaST, and OptiBET across 359 acquisitions from GE, Philips, and Siemens scanners at 1.5T and 3.0T. Results revealed statistically significant effects ($p < 0.001$) for both vendor and field strength [38].

4.3.10 Implications for Alzheimer's Disease Research

The application of skull stripping in neurodegeneration research introduces domain-specific challenges. Brain atrophy—the hallmark of Alzheimer's disease—alters the anatomical relationships that skull stripping algorithms exploit [37, 48].

Recent research by Tinauer et al. analysed 990 matched ADNI images and discovered that skull stripping introduces shortcut learning. CNNs trained on skull-stripped images learned brain contours introduced through preprocessing rather than clinically relevant atrophy markers—a "Clever Hans effect" inflating apparent classification accuracy [48].

Novosad and Collins evaluated skull stripping on ADNI subjects and found that brain masks include more subarachnoid CSF in atrophied brains, failure to remove non-brain tissue causes over-estimation of cortical thickness, and poor skull stripping propagates errors to regional volume and atrophy estimates [37].

It is also important to note that variability of anatomy, age, and extent of brain atrophy impacts skull stripping for volumetric Alzheimer's analysis [40].

4.3.11 Strengths and Weaknesses of Skull Stripping Approaches

Methodology	**Strengths**	**Weaknesses**
Mathematical Morphology	Simple to implement; fast [41]	Parameter-dependent; noise-sensitive; risk of over/under-segmentation [3]
Intensity-Based	Uses fundamental image properties	Sensitive to noise, bias field, threshold choice; watershed over-segmentation [3]
Deformable Surface	Accurate and robust; boundary-aware [42]	Sensitive to initialization; noise-sensitive; may fail in non-standard cases [38]
Atlas-Based	Leverages anatomical priors; good when intensities are unreliable [39]	Dependent on registration quality; computationally intensive [37]
Hybrid	Combines strengths of multiple methods; often fully automatic [40]	Complex; inherits weaknesses of contributing methods [3]
Deep Learning	State-of-the-art performance; learns features automatically [44, 46]	Requires large datasets; expensive; black-box behavior [4]

TABLE 4.1: Comparative strengths and weaknesses of skull-stripping methodologies.

4.3.12 Primary Challenges

Major difficulties arise from [3, 35, 37]:

- MRI artifacts (noise, intensity bias, motion)
- Anatomical complexity and overlapping tissue intensities
- Presence of pathology (e.g., tumors) affecting segmentation accuracy
- Brain atrophy in neurodegeneration creating ambiguous tissue boundaries

4.4 Voxel-Based Morphometry (VBM)

VBM is a whole-brain statistical technique that replaced the manual procedure in which clinicians had to track a specific brain structure, such as the hippocampus, across sequential slices. This technique frees the user from being confined to a single region of interest and instead enables analysis of the entire brain, detecting even subtle structural changes.

The VBM pipeline consists of several key steps:

1. **Image Acquisition**
   This typically involves acquiring a high-resolution T1-weighted MRI scan.
2. **Spatial Alignment (Normalization)**
   All images must be aligned to a common template to allow voxel-wise statistical comparison.
3. **Tissue Segmentation**
   Tissues are automatically classified into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). This relies on spatial priors and voxel intensities. Formally:
   $$P(\text{Tissue Class} \mid \text{Intensity}, \text{Location})$$
4. **Modulation**
   Modulation preserves the total amount of tissue after spatial transformations using the Jacobian determinant. The choice to use modulation depends on the research question:
   - Without modulation: analysis reflects *concentration* (e.g., GM density per unit volume).
   - With modulation: analysis reflects *absolute volume*.

Modulation may miss microscopic changes, and it is suggested that modulation should not be used to detect mesoscopic abnormalities such as cortical thinning [49].

5. **Smoothing**
   An isotropic Gaussian kernel is applied to:

   (a) Increase signal-to-noise ratio by averaging neighboring voxels.

   (b) Compensate for minor spatial misalignments after normalization.

   (c) Render the data more normally distributed for parametric statistics.

6. **Statistical Analysis**
   Voxel-wise statistical tests are typically performed using the General Linear Model (GLM). A t-value map (Statistical Parametric Map) is created by comparing group means and accounting for variance.

   To control false positives:

   - **Family-Wise Error (FWE)** correction
   - **False Discovery Rate (FDR)**

4.4.1 Evolution of VBM

VBM initially faced criticism regarding registration and segmentation accuracy. Two major improvements emerged:

- **Optimized VBM**
  Introduced study-specific templates and better skull-stripping to improve normalization accuracy.

- **DARTEL**
  A diffeomorphic algorithm using exponentiated Lie algebra for high-dimensional registration.

  The algorithm iteratively computes an increasingly accurate group average and stores subject-specific deformations as *individual flow fields*.

4.4.2 Alzheimer's Disease and Atrophy Patterns

1. **Medial Temporal Lobe**
   The earliest and most critical changes occur here, involving structures essential for episodic memory.

2. **Temporoparietal Association Cortex**
   Includes the temporal lobe, parietal lobe, and angular gyrus—responsible for memory, spatial cognition, and language processing.

3. **Posterior Cingulate Cortex and Precuneus**
   Key nodes of the Default Mode Network (DMN), implicated in self-referential thought and autobiographical memory.

4. **Frontal Lobe Involvement**
   VBM helped demonstrate that the frontal lobe is also affected, which has significant clinical implications for symptoms such as apathy and disinhibition.

5. **Unaffected Regions**
   The primary visual cortex, primary sensorimotor cortex, and the cerebellum remain largely intact.

6. **VBM and Braak Staging**
   Braak staging describes the sequential spread of neurofibrillary tangles (NFTs). VBM-detected atrophy patterns mirror those found in post-mortem studies, suggesting VBM can track macroscopic consequences of NFT accumulation.

7. **Structural Integrity vs. Metabolism**
   Hypometabolism may precede structural loss, meaning that VBM cannot always capture early dysfunction [50, 51].

4.4.3 VBM as a Prognostic Tool

Mild Cognitive Impairment (MCI) is heterogeneous. VBM can distinguish:

- **MCI converters**: likely to progress to AD
- **MCI non-converters**: may remain stable or revert

[52]

4.4.4 Differences Between Converters

MCI converters show atrophy patterns similar to mild AD, involving:

- Entorhinal cortex
- Hippocampus
- Temporal lobe

- Parietal lobe
- Posterior cingulate cortex

Non-converters show more limited and localized GM loss [53].

4.4.5 Meta-Analytic Power of VBM

A major meta-analysis identified the **left hippocampus and parahippocampal gyrus** as the most consistent predictors of conversion to AD [54]. Another found robust atrophy in the **left amygdala** and **right hippocampus**.

Rate of Decline in AD

Longitudinal VBM shows that "fast decliners" have greater GM loss at baseline, demonstrating that initial atrophy predicts future trajectory [55].

Clinical Utility

Tools such as VSRAD provide clinicians with voxel-wise z-scores derived from large normative datasets, similar to how blood test references are used.

Accuracy

VBM often achieves an AUC exceeding 0.90 in distinguishing AD from controls [56], demonstrating high diagnostic accuracy.

4.4.6 Subtypes of AD

VBM helps distinguish:

- **Early-Onset AD (EOAD)** – more parietal and posterior cingulate atrophy
- **Late-Onset AD (LOAD)** – more medial temporal atrophy

[57]

Structure–Clinical Correlations

VBM correlates brain atrophy with clinical features:

- Frontal atrophy correlates with impaired daily living activities.
- Apathy correlates with atrophy in anterior cingulate and orbitofrontal cortex.
- Disinhibition correlates with orbitofrontal atrophy.

4.4.7 Limitations of VBM

1. **Multiple Comparisons**
   Strict corrections (FWE) reduce false positives but may reduce sensitivity.

2. **Preprocessing Sensitivity**
   Results depend heavily on choices in normalization, segmentation, and statistical thresholds.

3. **Registration Problems**
   Even DARTEL may misalign regions due to individual variability.

4. **Segmentation Errors**
   Artifacts, low SNR, and partial volume effects can misclassify tissue.

5. **Ambiguity**
   A decrease in voxel intensity could reflect thinning, reduced surface area, or folding changes.

4.4.8 Machine Learning and VBM

1. Supervised methods (SVM, Random Forests) classify AD vs. controls and predict cognitive decline.

2. Feature importance can reveal novel biomarkers.

4.4.9 Deep Learning

CNNs can classify AD and MCI without hand-crafted features, achieving high accuracy (e.g., 80.9% for MCI identification) [58].

Chapter 5

Classification Techniques for Dementia Image Analysis

5.1 Foundational Machine Learning Paradigms in Dementia Classification

Before the widespread adoption of deep learning, classical machine learning pipelines followed a multi-step process consisting of:

1. feature extraction,
2. feature selection and dimensionality reduction,
3. classification using a machine learning model.

This pipeline formed the basis of early computational dementia research [59, 60].

5.1.1 Support Vector Machines (SVM)

Support Vector Machines (SVMs) were among the earliest and most widely used algorithms for binary classification. An SVM works by identifying an optimal hyperplane that separates data from different classes while maximizing the margin—the distance between the hyperplane and the nearest data points. For non-linearly separable data, SVMs employ the *kernel trick*, using polynomial or radial basis function (RBF) kernels to implicitly map data into a higher-dimensional space where linear separation becomes possible.

In dementia research, SVMs were extensively applied for classifying Alzheimer's disease (AD) versus healthy controls (HC). Input features were typically derived from struc-

tural MRI, often by dividing the brain into regions of interest (ROIs) and computing measurements such as grey matter (GM) volume, cortical thickness, or voxel-based morphometry (VBM) values.

Several studies reported high accuracies. For example, one study using whole-brain MRI-derived features achieved a mean classification accuracy of 94.5% (96.6% specificity, 91.5% sensitivity) for AD vs. controls [61]. Another reported an accuracy of 99.06% using 2D MRI slices [62].

However, performance degraded significantly when distinguishing more subtle clinical categories, such as healthy controls (HC), mild cognitive impairment (MCI), and in particular predicting MCI converters (MCI-C) versus non-converters (MCI-NC) [63].

The limitation lies not in the SVM algorithm itself, but in the nature of the disease and the constraints of manual feature engineering. Structural changes in late-stage AD—such as GM atrophy—are large and consistent across subjects, making the classes easily separable. MCI, however, is heterogeneous: not all MCI patients convert to AD, and subtle pathological differences are often not captured by handcrafted features such as ROI-based volumes.

This limitation highlighted the dependence of classical machine learning on high-quality engineered features, motivating the shift toward deep learning, where feature extraction is learned directly from the data.

5.1.2 Ensemble Methods: Decision Trees and Random Forests

Random Forests (RFs) emerged as a strong alternative to SVMs. An RF builds an ensemble of decision trees, each trained on a bootstrap sample of the data. At each node, splits consider only a subset of features, which reduces overfitting and is well-suited for high-dimensional, low-sample-size neuroimaging data [64].

The stability of RFs has been demonstrated in studies showing that even after substantial feature reduction, RF accuracy decreases less and remains more stable compared to multilayer perceptrons (MLPs) and convolutional neural networks (CNNs) [65].

A key advantage of RFs is their ability to compute feature importance via decreases in Gini impurity. This has repeatedly highlighted brain regions such as the hippocampus, amygdala, and inferior lateral ventricles as major contributors in NC/MCI/AD classification tasks [66].

RFs have achieved strong performance, including:

- 93.6% accuracy for predicting MCI-to-AD conversion from clinical data [67],

- 90.2% accuracy in three-class MRI-based classification (NC, MCI, AD) [66].

However, Gini-based importance scores may be biased toward continuous or high-cardinality features and can be misleading in the presence of correlated predictors [68]. In neuroanatomy, many structures (e.g., left/right hippocampus) are strongly correlated, meaning RF importance rankings require careful interpretation.

5.1.3 The Curse of Dimensionality: PCA and LDA

A major challenge in neuroimaging machine learning is the *curse of dimensionality* [69]. When the number of features p greatly exceeds the number of subjects, models tend to overfit and generalize poorly. Dimensionality reduction methods help mitigate this issue.

Principal Component Analysis (PCA)

PCA is an unsupervised method that identifies orthonormal components capturing directions of maximal variance [69, 70]. Retaining only the top components reduces noise and redundancy [71]. PCA has been used to compress VBM maps and ROI volumes before feeding them into classifiers, often improving performance.

Linear Discriminant Analysis (LDA)

LDA is a supervised method that seeks projections maximizing the ratio of between-class to within-class variance [72]. Unlike PCA, which is label-agnostic, LDA explicitly optimizes separability among classes such as NC, MCI, and AD.

Studies consistently show LDA outperforming PCA for dementia classification [69]. However, as a supervised method, it may overfit if distributions shift between train and test sets.

5.2 The Deep Learning Revolution

Deep learning disrupted the traditional feature-engineering workflow by enabling end-to-end learning from raw or minimally processed images [73].

5.2.1 Convolutional Neural Networks (CNNs)

CNNs form the backbone of most imaging-based dementia classification systems [74]. Their strength lies in hierarchical feature extraction: early layers detect low-level patterns (edges, textures), while deeper layers capture structural patterns associated with neurodegeneration [75].

2D vs. 3D CNN Designs

Choosing between 2D and 3D architectures involves trade-offs:

2D CNNs. These operate on slice-based inputs, enabling computational efficiency and transfer learning from natural-image models such as VGG, ResNet, Inception, and DenseNet [76, 77]. However, slice-wise inputs discard 3D anatomical continuity [78].

3D CNNs. 3D CNNs preserve volumetric structure using 3D kernels [79]. They achieve strong performance on AD/NC and multi-class tasks but require large datasets and memory, making them susceptible to overfitting [80].

Hybrid approaches include:

- 2.5D slices (central slice plus neighbors) [81],
- CNN+RNN architectures for sequential slice modeling [82].

Multimodal Fusion

Structural MRI (T1) and FDG-PET offer complementary structural and metabolic information. CNNs can fuse these modalities using multi-stream architectures, where early layers learn modality-specific features before being merged for joint classification [83, 84].

Interpretability: Grad-CAM

Deep models face adoption challenges due to limited interpretability. Grad-CAM provides visual explanations by generating heatmaps indicating which regions contributed most to the model's decision [5, 85].

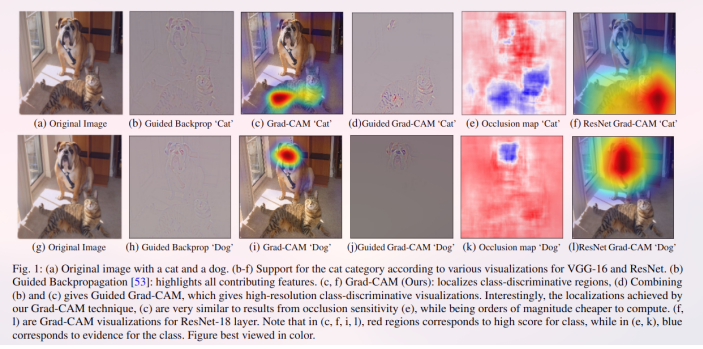

Fig. 1: (a) Original image with a cat and a dog. (b-f) Support for the cat category according to various visualizations for VGG-16 and ResNet. (b) Guided Backpropagation [53]: highlights all contributing features. (c, f) Grad-CAM (Ours): localizes class-discriminative regions, (d) Combining (b) and (c) gives Guided Grad-CAM, which gives high-resolution class-discriminative visualizations. Interestingly, the localizations achieved by our Grad-CAM technique, (c) are very similar to results from occlusion sensitivity (e), while being orders of magnitude cheaper to compute. (f, l) are Grad-CAM visualizations for ResNet-18 layer. Note that in (c, f, i, l), red regions corresponds to high score for class, while in (e, k), blue corresponds to evidence for the class. Figure best viewed in color.

FIGURE 5.1: Example Grad-CAM heatmap highlighting salient regions [5].

Such maps enable researchers and clinicians to verify that models rely on plausible neuroanatomical biomarkers rather than noise or artifacts.

5.2.2 Generative Models: GANs

Deep learning is limited by the scarcity of large, annotated medical datasets [84, 86]. Generative Adversarial Networks (GANs) address this by producing synthetic yet realistic MRI or PET images [87].

GAN-generated images can augment training datasets, alleviating class imbalance and improving classifier performance, with gains up to 11.68% reported [87]. GANs are also explored for:

- resolution enhancement (e.g., 1.5T $\rightarrow$ 3T MRI) [88],
- cross-modality synthesis (e.g., MRI-to-PET).

5.2.3 Vision Transformers (ViTs)

Transformers, originally developed for NLP, now challenge CNN dominance in medical imaging [89]. ViTs model long-range spatial dependencies using self-attention, enabling global context reasoning from the first layer [90]. This is appealing for dementia, where pathology is subtle and spatially distributed.

ViTs treat an image as a sequence of patches, which are embedded, combined with positional encodings, and processed through transformer layers [89]. Studies show ViTs achieving state-of-the-art AD classification performance [91].

5.3 Hybrid Classification Approaches

Hybrid methods combine CNNs for feature extraction with classical machine learning classifiers such as SVMs. These models leverage deep feature representations while benefiting from the robustness and theoretical properties of traditional classifiers. Such approaches have achieved up to 90% accuracy for AD vs. HC and up to 98% in four-class dementia classification tasks [92].

Chapter 6

Classification Performance and Accuracy Metrics

6.1 Evaluation Metrics

6.1.1 Accuracy, Sensitivity, Specificity

Evaluation metrics allow the visualization of the performance of a classifier and the comparison between different classifiers.

In Alzheimer's disease research, imbalanced data are extremely common. Imbalanced data occur when one class is significantly overrepresented or underrepresented.

To understand these metrics, we introduce the **confusion matrix**, which visualizes the relationship between predicted labels and actual labels:

Table 6.1: Confusion Matrix showing Actual vs. Predicted values

Total Population = P + N	**Predicted Positive**	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Accuracy in machine learning is the proportion of correct classifications over the total number of classifications. However, accuracy can be misleading, especially when dealing with imbalanced datasets. Its weaknesses include lower informativeness, discriminability, distinctiveness, and a bias toward the majority class [**?**].

Accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Sensitivity is the ability of a classifier to detect actual positives. In medical applications, a sensitivity of 0.8 means that the method correctly identifies positive cases 80% of the time.

$$\text{Sensitivity} = \text{TPR} = \frac{TP}{TP + FN}$$

Specificity measures the ability to correctly identify actual negatives:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

6.1.2 Precision and Recall

Precision (Positive Predictive Value) quantifies how many predicted positives are actually positive:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is equivalent to sensitivity:

$$\text{Recall} = \frac{TP}{TP + FN}$$

6.1.3 Model Validation Techniques

Model validation ensures that the model can generalize to unseen data. To improve robustness, methods such as **cross-validation** and **leave-one-out validation (LOOV)** are used.

- **Cross-validation**: Splits the dataset into k folds and trains the model k times, each time using one fold for testing and the rest for training.
- **Leave-One-Out Validation**: A special case where each data point is treated as a fold. Maximizes data usage but is computationally expensive.
- **Bootstrap Validation**: Repeatedly samples with replacement, trains on each sample, and tests on the remaining data. Effective for small or uncertain datasets [**?**].

6.1.4 ROC Curve

The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The Area Under the Curve (AUC) reflects classifier performance: the larger the AUC, the better the classifier.

AUC is robust to class imbalance [93, 94].

6.1.5 Matthews Correlation Coefficient

The Matthews Correlation Coefficient (MCC) is considered one of the most informative evaluation metrics [95]. Other related metrics include:

- Markedness
- Informedness (Bookmaker Informedness, BM)
- Balanced Accuracy (BA)

Chapter 7

Challenges in Dementia Image Analysis and Classification

7.1 Imbalanced Data

It is well understood from the literature that the performance of a machine learning (ML) model is upper bounded by the quality of the data [96]. Effective classification with imbalanced data is an important area of research, as high class imbalance is naturally inherent in many real-world applications, e.g., fraud detection and cancer detection [97].

In this problem, the majority of the sample data are labeled as one class and only a few samples are labeled as another class. This creates an imbalance in the dataset, and it is hard to solve because common classification algorithms are not designed to face these sorts of data [98]. In such a problem, almost all the examples are labeled as one class, while far fewer examples are labeled as the other class, usually the more important class. In this case, standard machine learning algorithms tend to be overwhelmed by the majority class and ignore the minority class since traditional classifiers seek accurate performance over a full range of instances [99]. In ML problems, differences in prior class probabilities—or class imbalances—have been reported to hinder the performance of standard classifiers such as decision trees [100].

When dealing with imbalanced datasets, standard evaluation metrics like accuracy can be misleading, as they are often biased toward the majority class. For a more robust and insightful assessment of classifier performance, it is crucial to employ metrics specifically designed to handle skewed class distributions. Based on [101], these can be divided into two primary categories: **Threshold Metrics** and **Ranking Methods**.

7.1.1 Threshold Metrics

Threshold metrics evaluate a classifier's performance at a fixed operational point. They provide a snapshot of performance but assume that the class distribution and error costs are known and constant.

Sensitivity and Specificity These metrics evaluate performance on each class independently.

- **Sensitivity (Recall, True Positive Rate):** Proportion of actual positive instances correctly identified.
- **Specificity:** Proportion of actual negative instances correctly identified.

Precision and Recall Widely used in information retrieval, these focus on the positive class.

- **Precision:** Proportion of predicted positives that are true positives.
- **Recall:** Equivalent to sensitivity.

Combined Metrics To simplify evaluation, the following metrics combine sensitivity, specificity, precision, and recall:

- **G-Mean:** Geometric mean of sensitivity and specificity, balancing performance across classes.
- **F-Measure:** Weighted harmonic mean of precision and recall, widely used in text categorization.

7.1.2 Ranking Methods and Metrics

Ranking methods evaluate performance across all decision thresholds, useful when class distribution or error costs vary.

ROC Analysis Plots True Positive Rate vs. False Positive Rate across thresholds. A curve near the top-left corner indicates a strong classifier. Particularly effective for imbalanced data.

AUC Area Under the ROC Curve; represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative one.

Precision–Recall Curves Show precision–recall trade-offs across thresholds and are preferred for highly imbalanced datasets.

In conclusion, evaluating classifiers on imbalanced datasets requires combining threshold metrics with ranking-based metrics. While threshold metrics evaluate specific operational conditions, ranking methods such as ROC and PR curves provide a more holistic understanding of classifier performance.

7.2 Limited Datasets

Big data and labels are not always available. Often we only have limited labeled data, such as medical images requiring expert annotation. ML with limited data poses major challenges. Domain shift is a critical issue in few-shot learning when training and testing domains differ [102]. Few datasets with limited data exist to train ML models [103], and robust models are hindered by the lack of large annotated datasets [104].

A recent study reports a dementia progression ML model built using a "limited, real-world clinical dataset," demonstrating that dataset scarcity harms model generalizability and clinical applicability [105]. Similarly, a comprehensive review highlights technical barriers and challenges in ML applications for dementia, particularly data scarcity and dataset diversity constraints [106].

7.2.1 Context and Challenges

Recent research consistently identifies the lack of large-scale, heterogeneous datasets (neuroimaging, biomarkers, clinical records) as a bottleneck for accurate and generalizable dementia ML models. Data limitations affect early diagnosis, subtype prediction, and progression tracking.

7.3 Small Sample Sizes

The combination of DL image analysis and large-scale imaging datasets presents major opportunities, yet barriers such as data availability, interpretability, and logistical constraints persist [107]. Further challenges include generalization, lack of gold standards, privacy concerns, and safety [108]. Studies emphasize the scarcity of sample-size determination methodologies for ML in medical imaging [109]. A meta-analysis shows that smaller studies often report inflated diagnostic accuracy, indicating systematic bias [110].

7.4 Image Quality and Variability

MRI intensities are acquired in arbitrary units, making scans non-comparable across sites. Even after normalization, inter-scan variability remains due to scanner effects and technical artifacts [111]. Variance in volumetric measures can differ by up to 10× depending on acquisition conditions, influenced by scanner hardware and tissue contrast differences [112]. Additional work quantifies scanner make, model, and protocol variability in image quality, which limits generalizability of CNNs across sites [113, 114].

7.5 Interpretability in Medical Applications

ML models in medicine must be interpretable, shareable, reproducible, and accountable. Black-box models are often accurate but face regulatory resistance. Interpretable approaches (kernel methods, prototype-based learning, deep kernel models) can provide clinically meaningful insights. Other key components include fairness, uncertainty quantification, and privacy-preserving collaborative learning such as federated learning and synthetic data via diffusion models [115]. Explainable AI enhances transparency and ethical compliance, though challenges remain in workflow integration and regulatory alignment [116]. ML has transformed healthcare across applications, though ethical concerns such as bias, privacy, and legal issues remain [117]. Challenges also persist in skin cancer detection due to data imbalance, interpretability limits, and model generalization difficulties [118].

7.6 Computational Costs

Deep learning progress correlates strongly with rising compute demands, which are becoming economically and environmentally unsustainable [119]. Efficient deep learning research explores model compression, optimization, and hardware acceleration. Compression techniques such as pruning, quantization, and knowledge distillation can reduce model size by over 100× while maintaining accuracy [120]. Lightweight models are essential for deployment on resource-constrained devices [121]. Architectures such as MobileNet and EfficientNet address this need while sustaining accuracy [122]. Transfer learning reduces training costs, improving performance with limited data [123]. Pruning and knowledge distillation remain critical for shrinking large models [124–126]. Additional work applies these techniques to Alzheimer's MRI classification [127].

7.7 Data Distillation

Data distillation provides a promising avenue for efficient and secure sharing of medical imaging datasets. Instead of sharing full datasets, distilled representations can preserve modeling performance while reducing data volume. Recent investigations show that a small, representative set of distilled images can achieve near-equivalent model performance, demonstrating potential for scalable clinical collaboration [128].

Chapter 8

Recent Advances And Innovations

8.1 Multi-Modal Integration

Most of the current research on Alzheimer's Disease (AD) classification relies on single modality data. A modality, as defined in [**?**], is an experience like sound, image, or touch.

Even though the scope of this work is not to review or examine works in data fusion, in order to provide a clearer understanding we should mention the challenges that are provided in one of the recent taxonomies of the field. The main challenges are:

1. Representation
2. Translation
3. Alignment
4. Fusion
5. Co-Learning

In one study, they integrated modalities from MRI, genetic (Single Nucleotide Polymorphisms (SNP)), and clinical test data to build a deep learning network by fusing these. The network was composed of stacked denoising auto-encoders to extract features from clinical and genetic data, and a 3D Convolutional Neural Network (CNN) for imaging data. Using data from the ADNI dataset, they demonstrated that deep models outperformed shallow models (SVM, Decision Trees, Random Forests, KNN). In addition, they demonstrated that multi-modality outperforms single modality in metrics such as accuracy, precision, recall, and F1 score. The models also identified

the hippocampus and other brain areas as distinguishing features, consistent with AD literature [**?**].

Another study tried to predict MCI from AD using a deep learning approach of a multimodal recurrent neural network. The modalities were composed of cross-sectional neuroimaging biomarkers, longitudinal cerebrospinal fluid biomarkers, and cognitive performance biomarkers also obtained by the ADNI dataset. The results showed a significant performance boost of about 6% in accuracy (75% $\rightarrow$ 81%), while the authors also note that multi-modal approaches seem promising in predicting the MCI stage regarding who will benefit the most from a clinical trial or as a stratification approach within clinical trials [**?**].

Another study reported that its deep learning fusion network approach performed better overall in classifying NC, MCI, AD, and nADD across the range of clinical tasks [**?**].

From the above evidence, it is well understood that more and more research is being implemented in the field by integrating multiple modalities in order to achieve higher classification accuracy, since it seems likely that multi-modal data are more robust and provide a more holistic view of the disease.

8.2 Explainable AI

The role of explainability has become increasingly important in recent years, as deep learning achieves state-of-the-art results and exceeds previous approaches in domains where decision-making is high-stakes (e.g., healthcare, finance).

There are **three main pillars** that Explainable AI (XAI) aims to cover:

1. **Trust:** Allowing the end user to trust and understand the fundamental reasoning behind the decision-making.

2. **Reducing Bias & Increasing Fairness:** In models that produce wrong behaviors—either due to the overrepresentation of certain groups in human-curated data or algorithmic mistakes—the goal is to increase fairness and reduce bias.

3. **Model Improvement & Debugging:** Enabling developers to identify errors in the model's logic and improve performance.

8.3 Definitions and Taxonomy

In the field, there is no clear consensus regarding the definitions of certain terms. The two most debated definitions are **interpretability** and **explainability**. Even though there is a debate surrounding these terms, we will adopt a practical definition to move forward:

- **Interpretability** is considered the ability to understand the model through its parameters or a simple graph. The model can be easily understood by a human (intrinsic explainability).
- **Explainability** is considered the process of providing a valid explanation for the reasoning the model used to arrive at a specific decision (post-hoc explainability).

The **taxonomy** of techniques can be described as:

- **Local & Global:** (Scope of the explanation)
- **Perturbation & Gradient-Based Methods:** (Mathematical approach)
- **Model-Agnostic & Model-Specific:** (Applicability across different model architectures)

For a more concise approach, the most noteworthy techniques are **LIME** and **SHAP**. Both are primarily considered local approaches; however, SHAP can also be considered a global technique because it produces a sum of all feature attributions, allowing for a holistic view of the dataset.

8.3.1 LIME

In LIME, the goal is to produce a function g that approximates the output of the neural network f. It utilizes a weighting variable that is higher for small perturbations (where we want high fidelity to the original input) and lower for distant, random samples. The algorithm aims to minimize both the error of the surrogate model and its complexity. Therefore, we define the explanation as the argument that minimizes the following objective:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{8.1}$$

Where $\mathcal{L}$ represents the **Loss** (fidelity of the surrogate model) and Ω represents the **Complexity** of the explanation model.

8.3.2 SHAP

This technique borrows characteristics from Game Theory (specifically Shapley Values). It attempts to assign a score to each feature based on how much it "cooperates" with other features to change the output. The mathematical formulation for the explanation model g is described as:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j \cdot z'_j \tag{8.2}$$

Where z' is the simplified feature (coalition vector) and ϕ is the attribution score assigned to that feature. By aggregating the scores across different examples, SHAP can provide a holistic view of the importance of each feature in the network.

8.3.3 Other Approaches

Other memorable approaches include Saliency Maps (which are gradient-based) and Occlusion Sensitivity (which is perturbation-based).

For the final category involving concept-based explanations, a notable method is **TCAV** (Testing with Concept Activation Vectors).

8.4 Sanity Checks

Even though we can create explanations for a problem, we must be aware of the dangers and understand the capabilities of each approach. For this reason, it is suggested to perform sanity checks to ensure robustness. The sanity checks should include:

1. **Stability:** Models should provide the same (or very similar) explanation for small perturbations to the input, up to a constant ϵ.

2. **Robustness:** Models injected with adversarial examples should allow the explanation method to identify why those examples are widely different from the distribution.

3. **Determinism:** For the exact same input, models should strictly produce the same explanation.

8.5 Related Work

In this work, the authors tried to create a network by combining a predictor with an explainable tool in order to provide accurate diagnosis while using visualization maps to confirm prediction basis. They built a predictor based on an attention mechanism using multi-scale features to teach a network to predict the correct labels representing the input features. They state that the network shows state-of-the-art accuracy and explainability and is able to define critical areas more clearly and with less noise that matches the neuroscience background literature [**?**].

In another study, researchers used data-level fusion of clinical data, MRI segmentation data, and psychological data. They employed many algorithms with Random Forest being the best and achieving the highest score value. They also used SHAP explanations for further explainability [**?**].

Additionally, in a systematic review on explainable AI in Alzheimer's research, it can be seen that most of the approaches contain post-hoc and model-agnostic approaches. Techniques such as SHAP, LIME, Grad-CAM, and LRP are shown to dominate the field. Also interesting are novel approaches using other modalities (speech & text), along with the current trade-off between accuracy and explainability in the domain of XAI [**?**].

8.6 Data Augmentation & Transfer Learning

Even though Deep Learning has performed tremendously in many computer vision tasks and has become a methodology of choice for analyzing medical images [**?**], these large networks usually require vast amounts of data in order to avoid overfitting.

Overfitting refers to the process where a model learns a function f with very high variance, perfectly modelling the training data and not being able to generalize to new instances that do not belong to the prior distribution of the training data. In many applications, such as medical imaging, data is limited due to a multitude of factors, but mainly due to the scarcity of datasets annotated by specialists, which is labor-intensive and has a high cost.

Data Augmentation is a collection of techniques used to increase both the quality and the size of an existing dataset to build better deep learning models. The image augmentation algorithms include the following [6] (See Table 8.1 and Figure 8.1).

Table 8.1: Image Augmentation Techniques

Technique	Description
1. Geometric transformations	This category encompasses simple transformations such as flipping, cropping, rotation, and translation. These methods are generally easy to implement and effectively address positional biases in the training data. However, practitioners must ensure transformations are "safe" and preserve the label (e.g., rotating a '6' might turn it into a '9').
2. Color space augmentations	Also called photometric transformations, these techniques manipulate pixel values in the RGB color channels or image histograms. Common methods include adjusting brightness, contrast, or isolating color channels to help models overcome lighting biases present in testing data.
3. Kernel filters	This technique involves sliding an $n \times n$ matrix across an image to sharpen or blur it. Blurring can improve resistance to motion blur, while sharpening emphasizes details. A specific variation called PatchShuffle Regularization randomly swaps pixel values within a window to improve robust feature learning.
4. Mixing images	This counterintuitive approach mixes multiple images to create new training instances. Methods include SamplePairing (averaging pixel values of two images) or concatenating random crops. Despite producing unnatural-looking images, this method has proven effective at reducing error rates.
5. Random erasing	Inspired by dropout regularization, this technique randomly selects an $n \times m$ patch of an image and masks it with 0s, 255s, or random values. By removing parts of the input, it forces the model to pay attention to the entire image rather than just a subset, helping address occlusion challenges.
6. Feature space augmentation	Instead of manipulating input images, this technique operates on the lower-dimensional vector representations (feature space) found in high-level network layers. Techniques include adding noise to these vectors or performing interpolations between nearest neighbors (similar to SMOTE) to generate new instances.

TABLE 8.1: Image Augmentation Techniques (continued)

Technique	Description
7. Adversarial training	This involves using a rival framework to generate "adversarial attacks"—constrained noise injections that cause misclassifications. Using these adversarial examples during training acts as a search algorithm for augmentations, strengthening weak decision boundaries and improving model robustness.
8. GANs	Generative Adversarial Networks (GANs) use a "generator" network to create artificial images and a "discriminator" network to distinguish real from fake ones. This powerful oversampling technique can "unlock" additional information from a dataset and create synthetic training data to increase size and diversity.
9. Neural Style Transfer	This algorithm manipulates sequential representations in a CNN to transfer the artistic style or texture of one image to another while preserving content. It is particularly useful for randomizing environments (e.g., lighting and texture) when transferring models from simulations to the real world.
10. Meta-learning	This refers to using neural networks to optimize other neural networks, specifically for finding optimal augmentation strategies. Examples include "Neural Augmentation" (learning style transfer weights), "Smart Augmentation" (merging images via a network), and "AutoAugment" (using Reinforcement Learning to find optimal transformation policies).

FIGURE 8.1: Showcases all the available image augmentation techniques. Courtesy of [6].

Furthermore, distinct approaches have been developed to generate synthetic training data; one study utilized Deep Convolutional Generative Adversarial Networks (DCGANs) to synthesize Positron Emission Tomography (PET) images across three disease stages, effectively overcoming the lack of labeled data [**?**]. Similarly, other research employed data augmentation within a transfer learning framework to rectify severe class imbalance in 3D Magnetic Resonance Imaging (MRI) datasets from the OASIS dataset

[?]. In both instances, the application of these augmentation strategies resulted in significant improvements in model accuracy and diagnostic performance.

8.6.1 Transfer Learning

Transfer Learning aims as a technique to improve the performance of a learner in a targeted domain by leveraging knowledge of the learner in other domains that are related. By this method, the data of the target domain can be considerably decreased. In medical imaging, whereas it has been mentioned that data scarcity remains an issue, transfer learning seems promising to overcome this obstacle and produce state-of-the-art transfer learners and leverage domains that are similar with more data to construct better networks.

The different categorizations of transfer learning can be seen in Figure 8.2, provided by [?].

FIGURE 8.2: Categorizations of transfer learning [?].

Several approaches in Alzheimer's research leverage transfer learning to detect and classify Alzheimer's disease.

In one study, transfer learning was used to classify four stages of Alzheimer's Disease: Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented. The deep learning model made use of brain MRI scans and achieved an accuracy of 91.7%, outperforming previous approaches. More specifically, they used a modified AlexNet architecture which was trained on ImageNet datasets as a source domain to perform knowledge transfer [?].

Another study used a VGG architecture with already pre-trained weights. They used the foundational network to perform transfer learning and optimized the network using additional MRI images. They show through experiments that with a size almost 10 times smaller on the OASIS MRI dataset, they can perform comparable to or better than some current deep learning approaches [?].

Chapter 9

Gaps in Current Literature and Future Research Directions

9.1 Generalizability and Standardized Testing

Although AI is performing increasingly well in medical domains, it still lacks the generalizability testing required for wider adoption. This is due to researchers and clinicians having limited access to diverse data—different from the training data—which is needed to test model robustness and reliability. This article provides recommendations for creating "standardized tests" (benchmark datasets) to solve this problem [**?**].

Another article goes beyond validating models and proposes a framework that is standard across pre-processing and image acquisition protocols. In the case of Alzheimer's disease, this is even more important, as different scanning protocols hinder the ability of models to generalize or limit the amount of training data. The work focuses on radiomics, but its broader implications for medical imaging and computer-aided diagnosis are relevant to Alzheimer's and dementia [**?**].

Other studies in radiomics have demonstrated the impact that different acquisition protocols can have on diagnosis and model accuracy by extensively validating algorithms on diverse datasets. The findings highlight a significant decrease in performance when acquisition diversity is not properly addressed [**?**].

9.2 Interpretable Models

Taking influence from the pillars previously set for explainability, there is a need for better explanations of model decision-making, especially in domains that affect human life, such as medicine. The pillars discussed are three.

9.2.1 Trust and Clinical Adoption

The first pillar concerns the need to cultivate trust among patients and clinicians, who will be responsible for conducting computer-aided diagnosis. The ability to examine the decision-making process and access the model's reasoning will vastly enhance clinicians' trust in these systems. This, in turn, improves both the quality of diagnosis and the number of patients that clinicians can handle, while maintaining consistent performance across individuals.

9.2.2 Reducing Bias and Increasing Fairness

The second pillar focuses on reducing bias and improving fairness. As stated previously, models are constrained by the distribution of the data on which they are trained. The dataset defines the prior distribution and the upper limit of information captured by the model, even when accounting for generalization. If the data are skewed or do not represent a stratifiable distribution of the patient population, models may exhibit bias and lead to incorrect diagnoses. Explainable AI helps mitigate this by providing reasoning behind decisions for clinicians to assess.

9.2.3 Better Model Development and Debugging

The third pillar emphasizes improved model development and system debugging. Access to the model's reasoning allows developers to identify limitations and provide rigorous debugging to improve performance more broadly.

These are the main pillars that make explainable AI crucial for the further development of systems, trust with end-users, and the avoidance of bias in decision-making. This is why explainability and interpretability remain major challenges in deploying deep learning for disease progression or classification outside research environments.

9.3 Detection at Different Stages

Based on the revised biomarker criteria [**?**], the integration of MRI and PET biomarkers—or their fusion—is of immense importance. Multimodal fusion can serve as the backbone of Alzheimer's disease (AD) detection and enhance the ability to track disease progression, which is vital for developing drugs aimed at slowing or curing neurodegeneration.

AD is a complex pathology in which multiple factors contribute to neurodegeneration; early pathological signs, such as amyloid-beta ($A\beta_{42}$) deposition, can be observed up

to 20 years prior to clinical symptoms [**?**]. Nevertheless, **early prediction remains challenging due to subtle brain changes that are difficult to quantify** [**?**].

Deep learning and advanced methods—such as data augmentation, synthetic data generation, transfer learning, and multimodal fusion—can significantly enhance early disease classification, enable longitudinal patient monitoring, and help evaluate treatment efficacy.

Chapter 10

Limitations

While there are a lot of studies that showcase significant improvement in disease classification through MRI scans, it is necessary to accept that several limitations exist before any clinical utility can be recognized. These limitations arise from the inherent variability from multiple sources and the different acquisition protocols and systems, from the experiment architectures and the intricacies of different approaches, along with the heterogeneity of the disease. Understanding these limitations is essential for future research and building more robust, reliable and clinically translatable systems.

10.1 Data Leakage

The primary limitation that has been researched in the field of medical imaging and machine learning is that of data leakage, which presents a serious obstacle for clinical translation. As documented in [129], nearly half of the studies between 2017 and 2019 failed to separate same subject data from contaminating the testing set. This leakage creates artificially inflated accuracy metrics that do not reflect generalized diagnostic capability.

The most common cause of data leakage seems to happen through "slice-level" splitting, where MRI slices from the same patient are distributed across both sets (training, testing). This leads to model overfitting, a poor indicator of clinical performance and generalization or broader pattern recognition.

Pulido-Salalgado et al. (2025) [130] demonstrated this effect empirically within a single study, showing a 28-percentage-point accuracy drop when switching from slice-wise to subject-wise splitting. The magnitude of this drop—equivalent to the difference between a highly promising diagnostic tool and one barely exceeding chance—illustrates why methodological rigor is non-negotiable.

Furthermore, data contamination can persist in other forms. Longitudinal studies usually include data from the same subject at different disease stages, and if these visits are not carefully tracked, temporal correlations can produce data leakage. Similarly, datasets that aggregate data from multiple acquisition sites may include the same patient scanned at different facilities. In Young (2025) [130], it was found that only 4.5% of published studies implemented the complete "methodological triad" of subject-wise splitting, external validation, and confounder control, suggesting that the field's reported performance metrics are systematically optimistic.

10.2 The Accuracy Paradox and Class Imbalance

The dataset employed in this study, drawn from the Alzheimer's Disease Neuroimaging Initiative (ADNI), exhibits significant class imbalance typical of clinical populations. Healthy Controls (HC) and Mild Cognitive Impairment (MCI) cases substantially outnumber confirmed Alzheimer's Disease cases. This imbalance renders overall accuracy an unreliable and potentially misleading performance metric, a phenomenon termed the "accuracy paradox" by Dubray et al. (2024) [131].

The dataset employed in this study also exhibits significant class imbalance as can be seen in Figure 10.1.

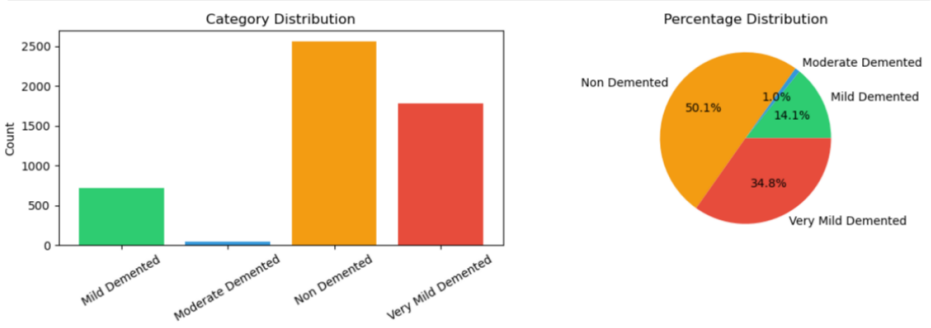

FIGURE 10.1: Class distribution showing the imbalance in the dataset.

This imbalance, as has already been mentioned, renders accuracy as an unreliable metric and more precise and robust metrics have been proposed (F1-score, AUC, MCC, etc.). The phenomenon of imbalance leading to misleading performance has been termed as "accuracy paradox" in Dubray et al. (2024) [131].

To frame this in practical terms, imagine a medical test that correctly identifies 9 out of 10 people, but only because it tells everyone they are healthy. The 1 in 10 patients with Alzheimer's Disease would receive false reassurance, potentially delaying critical interventions during the narrow window when treatments are most effective. This

scenario illustrates why clinical AI systems must be evaluated on metrics beyond raw accuracy.

For this reason, the results presented in this thesis prioritize sensitivity, specificity, F1-score, and Area Under the Receiver Operating Characteristic curve (AUC-ROC) alongside accuracy. We employed data augmentation using the MONAI (open source framework), weighted loss functions, and balanced sampling to address class imbalance during training. Moreover, additional frameworks like Grad-CAM can provide interpretability and were used in this study to assess model pattern recognition and be able to provide clinicians with an explanation of what the model evaluates as most important.

Nevertheless, residual bias toward the majority class likely persists. The model may still exhibit lower confidence and higher error rates when classifying true AD cases compared to healthy controls—a bias that would disproportionately affect the patients who most need accurate diagnosis.

10.3 Domain Shift and Generalizability

Most of the models where research is implemented are using curated datasets from research-grade MRI data (ADNI, OASIS) which employ standardized acquisition protocols across different sites fail when deployed on data acquired under different circumstances. While the standardized process reduces confounding variables during development, they introduce a critical limitation known as "domain shift."

10.3.1 Sources of Domain Shift in MRI Data

Magnetic field strength significantly impacts image resolution and for higher-field strength images contain a better signal-to-noise ratio, but can also amplify certain artifacts. A model thus that has been trained on specific images will fail to accurately predict correctly because of unfamiliar noise patterns or reduced visibility of subtle atrophy.

Acquisition sequence parameters—including echo time, repetition time, flip angle, and slice thickness—vary across institutions and protocols. What constitutes a "standard" T1-weighted sequence differs between research consortia and clinical departments. The model has learned one version of normal and may misinterpret technically valid but differently acquired images.

Diogo et al. (2022) [132] documented that models trained on research-grade images from one set frequently fail when applied to clinical-grade images from hospital settings. Dinsdale et al. (2024) [133] found that accuracy dropped to 71% when models were

tested on external datasets with different scanner configurations, compared to training data performance.

Critically, we did not perform external validation on a completely independent cohort (such as training on ADNI and testing on OASIS), meaning our reported metrics almost certainly overestimate real-world clinical performance.

10.4 Shortcut Learning and Region of Interest Bias

Another critical limitation is that deep learning models can learn shortcuts that are contained in medical images instead of actually learning the underlying conditions of the disease. This is called "shortcut learning." This is also mentioned in interpretability of deep neural networks as a way to ensure model understanding at the programming phase and also enhance clinician trust.

10.4.1 Skull-Stripping Artifacts

Preprocessing pipelines usually include a step for non-brain tissue (skull stripping) to focus the model directly on the brain tissue and its structure. However, as noted in Wen et al. (2020) [129], imperfect skull stripping can leave informative artifacts. If there is a difference in the stripping algorithm between healthy brains and atrophic brains—for example, leaving more residual scalp on atrophic brains where the boundary of brain-skull is less distinct—then it is likely that the model may learn to classify images based on preprocessing residuals rather than brain tissue characteristics.

The gap between brain surface and inner skull table increases with cortical atrophy, a hallmark of Alzheimer's Disease. If skull-stripping is incomplete, this enlarged cerebrospinal fluid space becomes visible as a classification cue. The model correctly identifies Alzheimer's cases, but for the wrong reason—it has learned to detect atrophy artifacts rather than analyze hippocampal or cortical tissue directly.

10.4.2 Biological Plausibility of Learned Features

In our experiments, we used Gradient-weighted Class Activation Mapping (Grad-CAM) for model interpretability, generating heatmaps indicating image regions that influenced the model's decision. However, it should be noted that interpretability methods are not validation methods, since a model can focus on ventricular enlargement which is a secondary issue but ignore the primary underlying pathology while still classifying a majority of samples correctly.

Young et al. (2025) [130] found that while 18.2% of studies used interpretability meth-

ods, only 12.5% of Grad-CAM implementations validated that highlighted regions corresponded to known neuropathological sites. This interpretability-validation chasm means that impressive heatmaps clustering around "reasonable" brain regions do not confirm that the model has learned disease-relevant features. Clinicians cannot trust explanations that lack rigorous validation against established pathological patterns. Figure 10.2 shows a Grad-CAM implementation along with the code.

```
from pytorch_grad_cam import GradCAM
from pytorch_grad_cam.utils.image import show_cam_on_image
from pytorch_grad_cam.utils.model_targets import ClassifierOutputTarget

def visualize_gradcam(model, img_tensor, true_label, target_layer):
    """Display GradCAM visualization for a single image."""
    img_tensor = img_tensor.to(device)

    # Create GradCAM object
    cam = GradCAM(model=model, target_layers=[target_layer])

    # Get prediction
    model.eval()
    with torch.no_grad():
        probs = F.softmax(model(img_tensor), dim=1)[0]
        pred = probs.argmax().item()

    # Generate GradCAM heatmap
    grayscale_cam = cam(input_tensor=img_tensor,
                        targets=[ClassifierOutputTarget(pred)])[0]

    # Prepare image for visualization
    img = img_tensor.cpu().squeeze().numpy()

    # Normalize image to [0, 1] range
    img_normalized = (img - img.min()) / (img.max() - img.min() + 1e-8)

    # Convert grayscale to RGB
    img_rgb = np.stack([img_normalized]*3, axis=-1).astype(np.float32)

    # Create overlay
    overlay = show_cam_on_image(img_rgb, grayscale_cam, use_rgb=True)
```

Listing 10.1: Grad-CAM visualization implementation

10.5 Ground Truth Uncertainty and Diagnostic Heterogeneity

Another important limitation is the reliability of diagnostic labels that are used for training. Machine learning models are upper-limited by the labels and data they are given, and it seems that clinical diagnosis of Alzheimer's Disease is imperfect. Wang et al. (2025) found that 71% of patients with dementia at autopsy had multiple coexisting pathologies, but only 67% had been clinically diagnosed with AD alone.

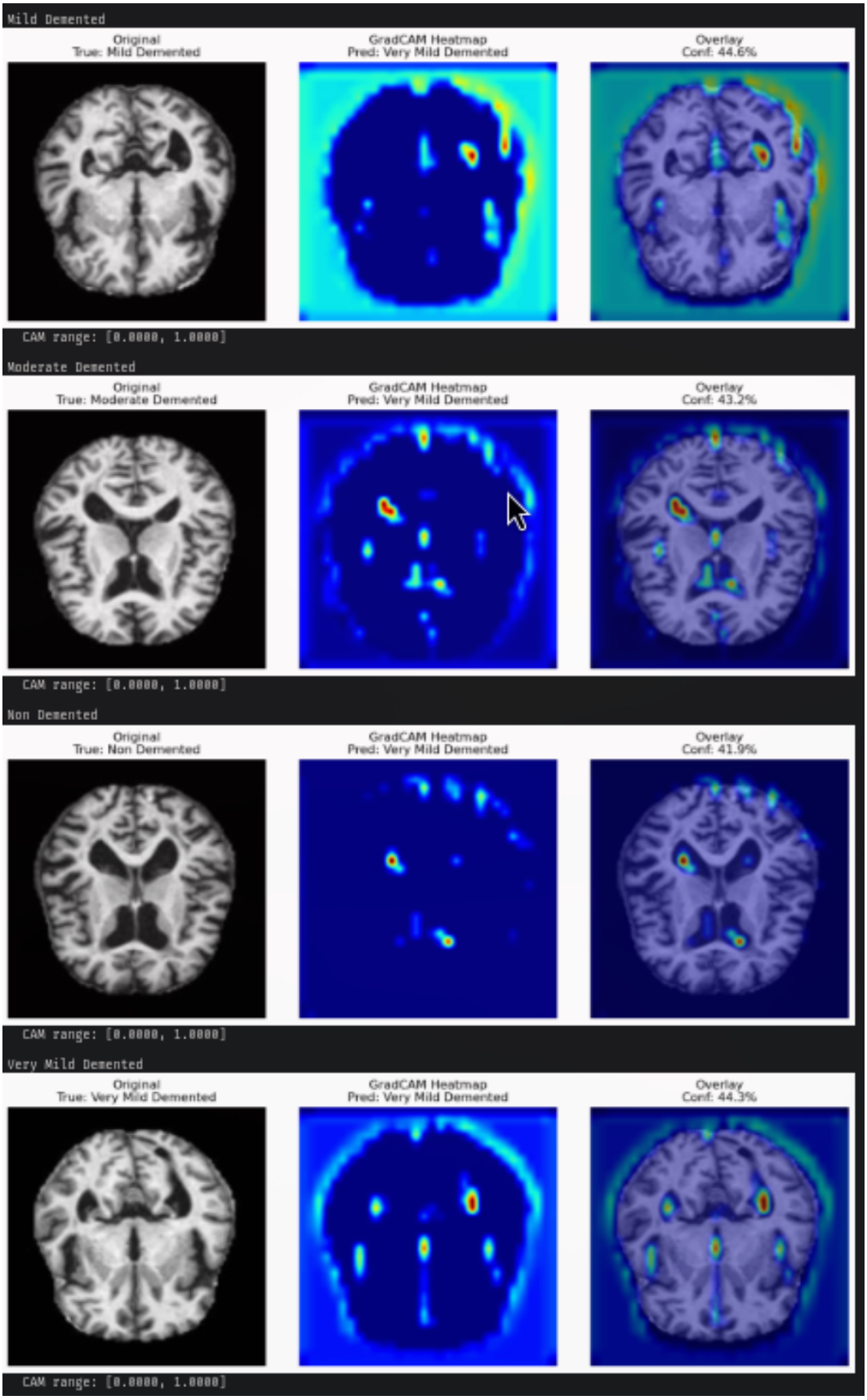

FIGURE 10.2: Grad-CAM visualization showing original image, heatmap, and overlay.

The same study demonstrated that when models were trained only on confirmed neuropathologically diagnoses instead of clinical labels, accuracy changed substantially—both for AD (84.4%) and vascular dementia (83.9%) but lower for Lewy body dementia (62.3%). This suggests that for different dementia subtypes there are different challenges for classification and that incorporating them under imprecise clinical labels degrades performance for all categories.

10.6 Disease Heterogeneity as an Intrinsic Limitation

Alzheimer's disease itself shows substantial phenotypic heterogeneity. In Young et al. (2018) [134], at least three distinct disease progression patterns were identified. Patients may present with predominantly hippocampal atrophy, predominantly cortical atrophy, or mixed patterns, with different sequences of regional involvement over time.

In another study, Kumar et al. (2024) [135] showed that even greater heterogeneity was found at more severe disease stages. Patients showcased distinct patterns of neurodegeneration, tau accumulation, and amyloid deposition. This presents a limitation where a binary classification between AD and control subjects might be inadequate since the disease manifests through multiple pathways with patient-specific trajectories.

10.7 Transparency, Reproducibility, and the Black Box Problem

Deep learning models operate as "black boxes," extracting features through millions of parameters without providing human-interpretable explanations for their decisions. Jo et al. (2019) emphasized that this opacity is particularly problematic for medical applications, where clinicians must justify diagnostic decisions to patients, families, and institutional review processes.

Reproducibility presents a related challenge. Dinsdale et al. (2024) [136] documented that none of the state-of-the-art studies they examined were fully reproducible. Authors rarely provide complete code, and critical implementation details—data augmentation strategies, exact model architectures, hyperparameter values, random seeds—are inconsistently reported. Sample sizes vary from 170 to 1,662 participants across studies, making cross-study benchmarking nearly impossible. Our own results, while documented as thoroughly as practical, may prove difficult to reproduce exactly due to inherent stochasticity in neural network training.

10.8 Limitations of Single-Modality Analysis

This study relies exclusively on structural MRI data, yet clinical diagnosis of AD integrates multiple information sources. In Leming et al. (2023) [137], it is mentioned that at least five modalities exist: structural imaging, functional imaging (PET, fMRI), cerebrospinal fluid biomarkers, genetic testing, and neuropsychological assessment. Relying solely on MRI can negatively impact information that can be leveraged for better model predictions.

However, even though the potential benefits of integrating multiple sources of information are documented [138, 139] and the combination of more modalities could improve accuracy, this can also introduce other complications. Data missingness, variable protocols, and dataset interoperability can increase model complexity and might compromise interpretability.

10.9 Summary of Limitations

The limitations presented in this chapter suggest that the metrics in our results represent upper bounds on clinical utility rather than deployment accuracy. Issues like data leakage may inflate accuracy. Class imbalance renders accuracy as misleading and masks reduced sensitivity for underrepresented classes. Domain shift implies that model performance will degrade if it encounters images from different scanners, protocols, or patient populations.

Shortcut learning means we cannot be confident that the model has learned disease-relevant features rather than confounding artifacts. Ground truth uncertainty implies that our training labels are themselves noisy, introducing irreducible error into the learning process. Disease heterogeneity suggests that the binary classification paradigm may be fundamentally mismatched to the biological reality of Alzheimer's Disease.

These limitations do not invalidate the work presented in this thesis, but they do circumscribe its claims. The model demonstrates proof-of-concept that convolutional neural networks can learn discriminative features from brain MRI data, but substantial additional work—external validation, prospective clinical trials, integration with existing diagnostic workflows—would be required before any clinical deployment could be contemplated. Future research should prioritize the "methodological triad" identified by Young et al. (2025): rigorous subject-level data splitting, external validation on independent cohorts, and systematic confounder control. Only through such rigorous methodology can the field move from promising laboratory results toward genuinely useful clinical tools.

Chapter 11

Conclusion

The field of research in Dementia and Alzheimer Classification is clearly defined by the Deep Learning Revolution. Most of the research has translated into integrating Deep Learning into multiple steps of the machine learning pipeline to maximize classification accuracy and significant gains in diagnosis. Deep Learning networks are used throughout the pipeline, like in pre-processing (Intensity Normalization, Registration , Skull Stripping , Volumetric Difference Estimation, Denoising) but also as a tool to classify between stages of neurodegeneration.

Additionally we have seen networks trained to learn a prior, like in denoising , to be turned into generative networks for image generation.

Moreover deep neural networks are classified as universal approximators , meaning that they can resemble a function given enough parameters. So based on the ability to show gains connected to the power equations of scaling laws we can conclude that the main bottleneck of the field is data scarcity.

As has been mentioned efforts in the field have generated datasets for research purposes but there is a lack of benchmark datasets to check model generalizability.

Moreover even though the research front can generate competent enough models to detect and classify different stages of dementia the lack of trust and explainability , behind the model's reasoning blocks the usage in clinical settings.

Although these problems exist , methods that can overcome the limitations of data scarcity such as image fusion or data augmentation and even synthetic data have been explored showing promising results.

Finally even though limitations exist , by accounting for the progress that has happened in the field and the ability of models to become better serving as a prior , along

with the fact that models perform state-of-the-art in research, it is highly likely that progress and future research will produce innovation to overcome data scarcity and model generalization as well as generate trust between clinicians , patients and Artificial Intelligence.

Appendix

Pathophysiology of Alzheimer

1. Amyloid-β (Aβ) pathology is generally thought to begin with the early deposition of Aβ42, a more aggregation-prone and fibrillogenic isoform of the peptide. As these deposits accumulate, they disrupt the surrounding neuronal environment and ultimately influence the stability and function of the tau protein network, which is essential for maintaining axonal structure and intracellular transport. The subsequent tau dysfunction and formation of neurofibrillary tangles contribute directly to progressive neuronal degeneration characteristic of Alzheimer's disease. Although the preferential production or accumulation of Aβ42 over other forms such as Aβ40 can be influenced by genetic factors, no single mechanism fully accounts for this shift. It is also important to note that many individuals produce Aβ42 without developing Alzheimer's disease; pathology emerges when the peptide accumulates beyond the brain's capacity for clearance, leading to plaque formation. [**?**]

2. Amyloid-β (Aβ) pathology is generally thought to begin with the early deposition of Aβ42, a more aggregation-prone and fibrillogenic isoform of the peptide. As these deposits accumulate, they disrupt the surrounding neuronal environment and ultimately influence the stability and function of the tau protein network, which is essential for maintaining axonal structure and intracellular transport. The subsequent tau dysfunction and formation of neurofibrillary tangles contribute directly to progressive neuronal degeneration characteristic of Alzheimer's disease. Although the preferential production or accumulation of Aβ42 over other forms such as Aβ40 can be influenced by genetic factors, no single mechanism fully accounts for this shift. It is also important to note that many individuals produce Aβ42 without developing Alzheimer's disease; pathology emerges when the peptide accumulates beyond the brain's capacity for clearance, leading to plaque formation. [**?**]

Bibliography

[1] H. Jung, "Basic Physical Principles and Clinical Applications of Computed Tomography," *Progress in Medical Physics*, vol. 32, pp. 1–17, Mar. 2021.

[2] E. Goceri, "Fully Automated and Adaptive Intensity Normalization Using Statistical Features for Brain MR Images," *Celal Bayar University Journal of Science*, vol. 14, pp. 125–134, Mar. 2018.

[3] P. Kalavathi and V. B. S. Prasath, "Methods on Skull Stripping of MRI Head Scan Images—a Review," *Journal of Digital Imaging*, vol. 29, pp. 365–379, June 2016.

[4] H. Z. U. Rehman, H. Hwang, and S. Lee, "Conventional and Deep Learning Methods for Skull Stripping in Brain MRI," *Applied Sciences*, vol. 10, p. 1773, Jan. 2020.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *International Journal of Computer Vision*, vol. 128, pp. 336–359, Feb. 2020.

[6] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, p. 60, Dec. 2019.

[7] F. Márquez and M. A. Yassa, "Neuroimaging Biomarkers for Alzheimer's Disease," *Molecular Neurodegeneration*, vol. 14, p. 21, June 2019.

[8] K. Kantarci and C. R. Jack, "Neuroimaging in Alzheimer disease: An evidence-based review," *Neuroimaging Clinics*, vol. 13, no. 2, pp. 197–209, 2003.

[9] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The Alzheimer's disease neuroimaging initiative," *Neuroimaging Clinics of North America*, vol. 15, no. 4, p. 869, 2005.

[10] "Alzheimer's Association 2025 Alzheimer's Disease Facts and Figures,"

[11] M. Fang, J. Hu, J. Weiss, D. S. Knopman, M. Albert, B. G. Windham, K. A. Walker, A. R. Sharrett, R. F. Gottesman, P. L. Lutsey, T. Mosley, E. Selvin, and J. Coresh, "Lifetime risk and projected burden of dementia," *Nature Medicine*, vol. 31, pp. 772–776, Mar. 2025.

[12] M. Prince, G.-C. Ali, M. Guerchet, A. M. Prina, E. Albanese, and Y.-T. Wu, "Recent global trends in the prevalence and incidence of dementia, and survival with dementia," *Alzheimer's Research & Therapy*, vol. 8, p. 23, July 2016.

[13] D. B. Plewes and W. Kucharczyk, "Physics of MRI: A primer," *Journal of Magnetic Resonance Imaging*, vol. 35, pp. 1038–1054, May 2012.

[14] G. B. Frisoni, N. C. Fox, C. R. Jack Jr, P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," *Nature reviews neurology*, vol. 6, no. 2, pp. 67–77, 2010.

[15] L. K. Ferreira and G. F. Busatto, "Neuroimaging in Alzheimer's disease: Current role in clinical practice and potential future applications," *Clinics*, vol. 66, pp. 19–24, Jan. 2011.

[16] S. Basu, T. C. Kwee, S. Surti, E. A. Akin, D. Yoo, and A. Alavi, "Fundamentals of PET and PET/CT imaging," *Annals of the New York Academy of Sciences*, vol. 1228, pp. 1–18, June 2011.

[17] A. Nordberg, J. O. Rinne, A. Kadir, and B. Långström, "The use of PET in Alzheimer disease," *Nature Reviews Neurology*, vol. 6, no. 2, pp. 78–87, 2010.

[18] X. Sun, L. Shi, Y. Luo, W. Yang, H. Li, P. Liang, K. Li, V. C. T. Mok, W. C. W. Chu, and D. Wang, "Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions," *Biomedical Engineering Online*, vol. 14, p. 73, July 2015.

[19] M. Shah, Y. Xiao, N. Subbanna, S. Francis, D. L. Arnold, D. L. Collins, and T. Arbel, "Evaluating intensity normalization on MRIs of human brain with multiple sclerosis," *Medical Image Analysis*, vol. 15, pp. 267–282, Apr. 2011.

[20] U. Bağcı, J. K. Udupa, and L. Bai, "The role of intensity standardization in medical image registration," *Pattern Recognition Letters*, vol. 31, pp. 315–323, Mar. 2010.

[21] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, and C. M. Crainiceanu, "Statistical normalization techniques for magnetic resonance imaging," *NeuroImage: Clinical*, vol. 6, pp. 9–19, Jan. 2014.

[22] L. G. Nyúl, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE transactions on medical imaging*, vol. 19, no. 2, pp. 143–150, 2000.

[23] L. G. Nyúl and J. K. Udupa, "Method for standardizing the MR image intensity scale," June 2003.

[24] P. Salome, F. Sforazzini, G. Brugnara, A. Kudak, M. Dostal, C. Herold-Mende, S. Heiland, J. Debus, A. Abdollahi, and M. Knoll, "MR Intensity Normalization Methods Impact Sequence Specific Radiomics Prognostic Model Performance in Primary and Recurrent High-Grade Glioma," *Cancers*, vol. 15, p. 965, Jan. 2023.

[25] M. Kociołek, M. Strzelecki, and R. Obuchowicz, "Does image normalization and intensity resolution impact texture classification?," *Computerized Medical Imaging and Graphics*, vol. 81, p. 101716, Apr. 2020.

[26] A. Pandey and A. Jain, "Comparative analysis of KNN algorithm using various normalization techniques," *International Journal of Computer Network and Information Security*, vol. 10, no. 11, p. 36, 2017.

[27] S. G. K. Patro and K. K. Sahu, "Normalization: A Preprocessing Stage," Mar. 2015.

[28] S. Albert, B. D. Wichtmann, W. Zhao, A. Maurer, J. Hesser, U. I. Attenberger, L. R. Schad, and F. G. Zöllner, "Comparison of Image Normalization Methods for Multi-Site Deep Learning," *Applied Sciences*, vol. 13, p. 8923, Jan. 2023.

[29] G. Collewet, M. Strzelecki, and F. Mariette, "Influence of MRI acquisition protocols and image intensity normalization methods on texture classification," *Magnetic Resonance Imaging*, vol. 22, pp. 81–91, Jan. 2004.

[30] F. Orlhac, J. J. Eertink, A.-S. Cottereau, J. M. Zijlstra, C. Thieblemont, M. Meignan, R. Boellaard, and I. Buvat, "A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies," *Journal of Nuclear Medicine*, vol. 63, pp. 172–179, Feb. 2022.

[31] U. Michelucci, "An Introduction to Autoencoders," Jan. 2022.

[32] P.-L. Delisle, B. Anctil-Robitaille, C. Desrosiers, and H. Lombaert, "Realistic image normalization for multi-Domain segmentation," *Medical Image Analysis*, vol. 74, p. 102191, Dec. 2021.

[33] C. Xu, Y. Sun, Y. Zhang, T. Liu, X. Wang, D. Hu, S. Huang, J. Li, F. Zhang, and

G. Li, “Stain Normalization of Histopathological Images Based on Deep Learning: A Review,” *Diagnostics*, vol. 15, no. 8, p. 1032, 2025.

[34] M. Elad, B. Kawar, and G. Vaksman, “Image Denoising: The Deep Learning Revolution and Beyond—A Survey Paper,” *SIAM Journal on Imaging Sciences*, vol. 16, pp. 1594–1654, Sept. 2023.

[35] A. Fatima, A. R. Shahid, B. Raza, T. M. Madni, and U. I. Janjua, “State-of-the-Art Traditional to the Machine- and Deep-Learning-Based Skull Stripping Techniques, Models, and Algorithms,” *Journal of Digital Imaging*, vol. 33, pp. 1443–1464, Dec. 2020.

[36] F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl, “A hybrid approach to the skull stripping problem in MRI,” *NeuroImage*, vol. 22, pp. 1060–1075, July 2004.

[37] P. Novosad, V. Fonov, and D. L. Collins, “Accurate and robust segmentation of neuroanatomy in T1-weighted MRI by combining spatial priors with deep convolutional neural networks,” Feb. 2019.

[38] R. Souza, O. Lucena, J. Garrafa, D. Gobbi, M. Saluzzi, S. Appenzeller, L. Rittner, R. Frayne, and R. Lotufo, “An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement,” *NeuroImage*, vol. 170, pp. 482–494, Apr. 2018.

[39] S. F. Eskildsen, P. Coupé, V. Fonov, J. V. Manjón, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Østergaard, D. L. Collins, and Alzheimer’s Disease Neuroimaging Initiative, “BEaST: Brain extraction based on nonlocal segmentation technique,” *NeuroImage*, vol. 59, pp. 2362–2373, Feb. 2012.

[40] J. E. Iglesias, C.-Y. Liu, P. M. Thompson, and Z. Tu, “Robust brain extraction across datasets and comparison with publicly available methods,” *IEEE transactions on medical imaging*, vol. 30, pp. 1617–1634, Sept. 2011.

[41] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, “Magnetic resonance image tissue classification using a partial volume model,” *NeuroImage*, vol. 13, pp. 856–876, May 2001.

[42] S. M. Smith, “Fast robust automated brain extraction,” *Human Brain Mapping*, vol. 17, pp. 143–155, Nov. 2002.

[43] R. W. Cox, “AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages,” *Computers and Biomedical Research, an International Journal*, vol. 29, pp. 162–173, June 1996.

[44] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, M. Bendszus, K. H. Maier-Hein, and P. Kickingereder, "Automated brain extraction of multisequence MRI using artificial neural networks," *Human Brain Mapping*, vol. 40, pp. 4952–4964, Dec. 2019.

[45] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller, "Deep MRI brain extraction: A 3D convolutional neural network for skull stripping," *NeuroImage*, vol. 129, pp. 460–469, Apr. 2016.

[46] A. Hoopes, J. S. Mora, A. V. Dalca, B. Fischl, and M. Hoffmann, "SynthStrip: Skull-stripping for any brain image," *NeuroImage*, vol. 260, p. 119474, Oct. 2022.

[47] L. Fisch, S. Zumdick, C. Barkhau, D. Emden, J. Ernsting, R. Leenings, K. Sarink, N. R. Winter, B. Risse, U. Dannlowski, and T. Hahn, "Deepbet: Fast brain extraction of T1-weighted MRI using Convolutional Neural Networks," Aug. 2023.

[48] C. Tinauer, M. Sackl, R. Stollberger, R. Schmidt, S. Ropele, and C. Langkammer, "Skull-stripping induces shortcut learning in MRI-based Alzheimer's disease classification," *Insights into Imaging*, vol. 16, p. 283, Dec. 2025.

[49] J. Radua, E. J. Canales-Rodríguez, E. Pomarol-Clotet, and R. Salvador, "Validity of modulation and optimal settings for advanced voxel-based morphometry," *NeuroImage*, vol. 86, pp. 81–90, Feb. 2014.

[50] M. T. Duong, S. R. Das, P. Khandelwal, X. Lyu, L. Xie, E. McGrew, N. Dehghani, C. T. McMillan, E. B. Lee, L. M. Shaw, P. A. Yushkevich, D. A. Wolk, I. M. Nasrallah, and Alzheimer's Disease Neuroimaging Initiative, "Hypometabolic mismatch with atrophy and tau pathology in mixed Alzheimer's and Lewy body disease," *Brain*, vol. 148, pp. 1577–1587, May 2025.

[51] G. Chételat, B. Desgranges, B. Landeau, F. Mézenge, J. B. Poline, V. de la Sayette, F. Viader, F. Eustache, and J.-C. Baron, "Direct voxel-based comparison between grey matter hypometabolism and atrophy in Alzheimer's disease," *Brain*, vol. 131, pp. 60–71, Jan. 2008.

[52] M. Bozzali, M. Filippi, G. Magnani, M. Cercignani, M. Franceschi, E. Schiatti, S. Castiglioni, R. Mossini, M. Falautano, G. Scotti, G. Comi, and A. Falini, "The contribution of voxel-based morphometry in staging patients with mild cognitive impairment," *Neurology*, vol. 67, pp. 453–460, Aug. 2006.

[53] L. G. Apostolova and P. M. Thompson, "Mapping Progressive Brain Structural

Changes in Early Alzheimer's Disease and Mild Cognitive Impairment," *Neuropsychologia*, vol. 46, no. 6, pp. 1597–1612, 2008.

[54] L. K. Ferreira, B. S. Diniz, O. V. Forlenza, G. F. Busatto, and M. V. Zanetti, "Neurostructural predictors of Alzheimer's disease: A meta-analysis of VBM studies," *Neurobiology of Aging*, vol. 32, pp. 1733–1741, Oct. 2011.

[55] "VBM anticipates the rate of progression of Alzheimer disease | Neurology." https://www.neurology.org/doi/abs/10.1212/01.wnl.0000303960.01039.43.

[56] K. Ishii, T. Kawachi, H. Sasaki, A. K. Kono, T. Fukuda, Y. Kojima, and E. Mori, "Voxel-Based Morphometric Comparison Between Early- and Late-Onset Mild Alzheimer's Disease and Assessment of Diagnostic Performance of Z Score Images," *AJNR: American Journal of Neuroradiology*, vol. 26, pp. 333–340, Feb. 2005.

[57] K. Ishii, T. Kawachi, H. Sasaki, A. K. Kono, T. Fukuda, Y. Kojima, and E. Mori, "Voxel-Based Morphometric Comparison Between Early- and Late-Onset Mild Alzheimer's Disease and Assessment of Diagnostic Performance of Z Score Images," *AJNR: American Journal of Neuroradiology*, vol. 26, pp. 333–340, Feb. 2005.

[58] H. Huang, S. Zheng, Z. Yang, Y. Wu, Y. Li, J. Qiu, Y. Cheng, P. Lin, Y. Lin, J. Guan, D. J. Mikulis, T. Zhou, and R. Wu, "Voxel-based morphometry and a deep learning model for the diagnosis of early Alzheimer's disease based on cerebral gray matter changes," *Cerebral Cortex*, vol. 33, pp. 754–763, Feb. 2023.

[59] T. Jo, K. Nho, and A. J. Saykin, "Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data," *Frontiers in Aging Neuroscience*, vol. 11, p. 220, Aug. 2019.

[60] J. Islam and Y. Zhang, "Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," *Brain Informatics*, vol. 5, p. 2, May 2018.

[61] "(PDF) Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI," *ResearchGate*.

[62] P. Niranjan Kumar, "SVM-Based Classifier For Early Detection Of Alzheimer's Disease," *Educational Administration: Theory and Practice*, pp. 1120–1131, May 2024.

[63] E. Pellegrini, L. Ballerini, M. d. C. V. Hernandez, F. M. Chappell, V. González-Castro, D. Anblagan, S. Danso, S. Muñoz-Maniega, D. Job, and C. Pernet, "Ma-

chine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 10, pp. 519–535, 2018.

[64] A. Sarica, A. Cerasa, and A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review," *Frontiers in Aging Neuroscience*, vol. 9, Oct. 2017.

[65] S. I. Dimitriadis and D. Liparas, "How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: From Alzheimer's disease neuroimaging initiative (ADNI) database," *Neural Regeneration Research*, vol. 13, pp. 962–970, June 2018.

[66] M. Song, H. Jung, S. Lee, D. Kim, and M. Ahn, "Diagnostic Classification and Biomarker Identification of Alzheimer's Disease with Random Forest Algorithm," *Brain Sciences*, vol. 11, p. 453, Apr. 2021.

[67] M. Velazquez and Y. Lee, "Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects,"

[68] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, p. 25, Jan. 2007.

[69] "(PDF) A Comparative Study of PCA and LDA for Dimensionality Reduction in a 4-Way Classification Framework." https://www.researchgate.net/publication/378806266_A_Comparative_Study_of_PCA_an Way_Classification_Framework.

[70] L. Lazli, "Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development," *JMIRx Med*, vol. 6, p. e60866, Apr. 2025.

[71] V. A. Miller, S. Erlien, and J. Piersol, "Support vector machine classification of dimensionally reduced structural MRI images for dementia," June 2014.

[72] M. M.Dessouky, M. A. Elrashidy, and H. M. Abdelkader, "Selecting and Extracting Effective Features for Automated Diagnosis of Alzheimer's Disease," *International Journal of Computer Applications*, vol. 81, pp. 17–28, Nov. 2013.

[73] D. Ozkan, O. Katar, M. Ak, M. A. Al-Antari, N. Yasan Ak, O. Yildirim, H. S. Mir, R.-S. Tan, and U. Rajendra Acharya, "Deep Learning Techniques for Automated Dementia Diagnosis Using Neuroimaging Modalities: A Systematic Review," *IEEE Access*, vol. 12, pp. 127879–127902, 2024.

[74] A. Ebrahimi and S. Luo, "Convolutional neural networks for Alzheimer's disease detection on MRI images," *Journal of Medical Imaging*, vol. 8, p. 024503, Mar. 2021.

[75] A. Ebrahimi and S. Luo, "Convolutional neural networks for Alzheimer's disease detection on MRI images," *Journal of Medical Imaging*, vol. 8, p. 024503, Mar. 2021.

[76] V. S M., K. D., and N. V C., "Deep Learning-Driven Alzheimer's Disease Classification: Custom CNN and Pretrained Architectures for Accurate MRI Analysis," *Journal of Soft Computing Paradigm*, vol. 7, pp. 31–43, Mar. 2025.

[77] "(PDF) Improved Classification of Alzheimer's Disease With Convolutional Neural Networks," in *ResearchGate*.

[78] M. U. Ali, K. S. Kim, M. Khalid, M. Farrash, A. Zafar, and S. W. Lee, "Enhancing Alzheimer's disease diagnosis and staging: A multistage CNN framework using MRI," *Frontiers in Psychiatry*, vol. 15, June 2024.

[79] B. Khagi and G.-R. Kwon, "3D CNN Design for the Classification of Alzheimer's Disease Using Brain MRI and PET," *IEEE Access*, vol. 8, pp. 217830–217847, 2020.

[80] X. Xu, L. Lin, S. Sun, and S. Wu, "A review of the application of three-dimensional convolutional neural networks for the diagnosis of Alzheimer's disease using neuroimaging," *Reviews in the Neurosciences*, vol. 34, pp. 649–670, Aug. 2023.

[81] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, and T. A. D. N. Initiative, "Convolutional Neural Networks-Based MRI Image Analysis for the Alzheimer's Disease Prediction From Mild Cognitive Impairment," *Frontiers in Neuroscience*, vol. 12, Nov. 2018.

[82] A. Ebrahimi and S. Luo, "Convolutional neural networks for Alzheimer's disease detection on MRI images," *Journal of Medical Imaging*, vol. 8, p. 024503, Mar. 2021.

[83] Y. Huang, J. Xu, Y. Zhou, T. Tong, and X. Zhuang, "Diagnosis of Alzheimer's Disease via Multi-Modality 3D Convolutional Neural Network," *Frontiers in Neuroscience*, vol. 13, p. 509, May 2019.

[84] Z. Zhao, J. H. Chuah, K. W. Lai, C.-O. Chow, M. Gochoo, S. Dhanalakshmi, N. Wang, W. Bao, and X. Wu, "Conventional machine learning and deep learning

in Alzheimer's disease diagnosis using neuroimaging: A review," *Frontiers in Computational Neuroscience*, vol. 17, p. 1038636, Feb. 2023.

[85] M. Wang, "Interpretable 2D and 3D Convolutional Neural Networks for Alzheimer's Disease in Brain Scans,"

[86] G. Marcus, "Deep Learning: A Critical Appraisal,"

[87] F. Konidaris, T. Tagaris, M. Sdraka, and A. Stafylopatis, "Generative Adversarial Networks as an Advanced Data Augmentation Technique for MRI Data:," in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, (Prague, Czech Republic), pp. 48–59, SCITEPRESS - Science and Technology Publications, 2019.

[88] X. Zhou, S. Qiu, P. S. Joshi, C. Xue, R. J. Killiany, A. Z. Mian, S. P. Chin, R. Au, and V. B. Kolachalama, "Enhancing magnetic resonance imaging-driven Alzheimer's disease classification performance using generative adversarial learning," *Alzheimer's Research & Therapy*, vol. 13, p. 60, Mar. 2021.

[89] "(PDF) Vision Transformers in Medical Imaging: A Comprehensive Review of Advancements and Applications Across Multiple Diseases." https://www.researchgate.net/publication/390372350_Vision_Transformers_in_Medical_Im

[90] "Vision transformer architecture and applications in digital health: A tutorial and survey - PMC." https://pmc.ncbi.nlm.nih.gov/articles/PMC10333157/.

[91] P. T. Krishnan, P. Krishnadoss, M. Khandelwal, D. Gupta, A. Nihaal, and T. S. Kumar, "Enhancing brain tumor detection in MRI with a rotation invariant Vision Transformer," *Frontiers in Neuroinformatics*, vol. 18, June 2024.

[92] "Hybrid CNN-SVM for Alzheimer's Disease Classification from Structural MRI and the Alzheimer's Disease Neuroimaging Initiative (ADNI)," in *2018 International Conference on Biomedical Engineering, Machinery and Earth Science (BEMES 2018)*, Francis Academic Press, 2018.

[93] J. X. C. Ke, A. DhakshinaMurthy, R. B. George, and P. Branco, "The effect of resampling techniques on the performances of machine learning clinical risk prediction models in the setting of severe class imbalance: Development and internal validation in a retrospective cohort," *Discover Artificial Intelligence*, vol. 4, p. 91, Nov. 2024.

[94] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, p. 100994, May 2024.

[95] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, p. 6, Jan. 2020.

[96] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Sharma Mittal, and V. Munigala, "Overview and Importance of Data Quality for Machine Learning Tasks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (Virtual Event CA USA), pp. 3561–3562, ACM, Aug. 2020.

[97] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, p. 27, Dec. 2019.

[98] N. U. Niaz, K. N. Shahariar, and M. J. A. Patwary, "Class Imbalance Problems in Machine Learning: A Review of Methods And Future Challenges," in *Proceedings of the 2nd International Conference on Computing Advancements*, (Dhaka Bangladesh), pp. 485–490, ACM, Mar. 2022.

[99] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *2008 Fourth International Conference on Natural Computation*, vol. 4, pp. 192–201, IEEE, 2008.

[100] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study1," *Intelligent Data Analysis*, vol. 6, pp. 429–449, Nov. 2002.

[101] N. Japkowicz, "Assessment Metrics for Imbalanced Learning," in *Imbalanced Learning* (H. He and Y. Ma, eds.), pp. 187–206, Wiley, 1 ed., June 2013.

[102] F. Yao, "Machine learning with limited data," Jan. 2021.

[103] A. Merkin, R. Krishnamurthi, and O. N. Medvedev, "Machine learning, artificial intelligence and the prediction of dementia," *Current Opinion in Psychiatry*, vol. 35, p. 123, Mar. 2022.

[104] S. A. Martin, F. J. Townend, F. Barkhof, and J. H. Cole, "Interpretable machine learning for dementia: A systematic review," *Alzheimer's & Dementia*, vol. 19, pp. 2135–2149, May 2023.

[105] Y.-C. Huang, T.-C. Liu, and C.-J. Lu, "Establishing a machine learning dementia progression prediction model with multiple integrated data," *BMC Medical Research Methodology*, vol. 24, p. 288, Nov. 2024.

[106] Y. Wang, S. Liu, A. G. Spiteri, A. L. H. Huynh, C. Chu, C. L. Masters, B. Goudey, Y. Pan, and L. Jin, "Understanding machine learning applications in dementia

research and clinical practice: A review for biomedical scientists and clinicians," *Alzheimer's Research & Therapy*, vol. 16, p. 175, Aug. 2024.

[107] "Challenges for machine learning in clinical translation of big data imaging studies," *Neuron*, vol. 110, pp. 3866–3881, Dec. 2022.

[108] M. Aljuhani, A. Ashraf, and P. Edison, "Use of Artificial Intelligence in Imaging Dementia," *Cells*, vol. 13, p. 1965, Nov. 2024.

[109] "Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review," *Canadian Association of Radiologists Journal*, vol. 70, pp. 344–353, Nov. 2019.

[110] L. Lu, Q. S. Phua, S. Bacchi, R. Goh, A. K. Gupta, J. G. Kovoor, C. D. Ovenden, and M.-S. To, "Small Study Effects in Diagnostic Imaging Accuracy," *JAMA Network Open*, vol. 5, p. e2228776, Aug. 2022.

[111] T. L. S. Benzinger, T. Blazey, C. R. Jack, R. A. Koeppe, Y. Su, C. Xiong, M. E. Raichle, A. Z. Snyder, B. M. Ances, R. J. Bateman, N. J. Cairns, A. M. Fagan, A. Goate, D. S. Marcus, P. S. Aisen, J. J. Christensen, L. Ercole, R. C. Hornbeck, A. M. Farrar, P. Aldea, M. S. Jasielec, C. J. Owen, X. Xie, R. Mayeux, A. Brickman, E. McDade, W. Klunk, C. A. Mathis, J. Ringman, P. M. Thompson, B. Ghetti, A. J. Saykin, R. A. Sperling, K. A. Johnson, S. Salloway, S. Correia, P. R. Schofield, C. L. Masters, C. Rowe, V. L. Villemagne, R. Martins, S. Ourselin, M. N. Rossor, N. C. Fox, D. M. Cash, M. W. Weiner, D. M. Holtzman, V. D. Buckles, K. Moulder, and J. C. Morris, "Regional variability of imaging biomarkers in autosomal dominant Alzheimer's disease," *Proceedings of the National Academy of Sciences*, vol. 110, Nov. 2013.

[112] F. Kruggel, J. Turner, L. T. Muftuler, and A. D. N. Initiative, "Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort," *Neuroimage*, vol. 49, no. 3, pp. 2123–2133, 2010.

[113] I. Montero, S. Sotoudeh-Paima, E. Abadi, and E. Samei, "Intra- and inter-scanner CT variability and their impact on diagnostic tasks," *Proceedings of SPIE–the International Society for Optical Engineering*, vol. 13405, p. 134054C, Feb. 2025.

[114] S. Bhosekar, P. Singh, D. Garg, V. Ravi, and M. Diwakar, "A Review of Deep Learning-based Multi-modal Medical Image Fusion,"

[115] A. Begüm Bektaş and M. Gönen, "Machine Learning for Medicine Must Be Interpretable, Shareable, Reproducible and Accountable by Design," *arXiv e-prints*, pp. arXiv–2508, 2025.

[116] "A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges," *Informatics in Medicine Unlocked*, vol. 51, p. 101587, Jan. 2024.

[117] "Machine Learning in Healthcare: A Review of Current Applications and Future Trends." https://ieeexplore.ieee.org/document/10968281.

[118] "A Review on the Applications of Machine Learning and Deep Learning Techniques for Skin Cancer Detection." https://ieeexplore.ieee.org/document/10933289.

[119] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The Computational Limits of Deep Learning," July 2022.

[120] "(PDF) Efficient Deep Learning: A Survey of Model Compression and Optimization Techniques for Resource-Constrained Environments." https://www.researchgate.net/publication/394538538_Efficient_Deep_Learning_A_Survey_of Constrained_Environments, Aug. 2025.

[121] H.-I. Liu, M. Galindo, H. Xie, L.-K. Wong, H.-H. Shuai, Y.-H. Li, and W.-H. Cheng, "Lightweight Deep Learning for Resource-Constrained Environments: A Survey," Apr. 2024.

[122] K. B. Nampalle, P. Singh, U. V. Narayan, and B. Raman, "DeepMediX: A Deep Learning-Driven Resource-Efficient Medical Diagnosis Across the Spectrum," July 2023.

[123] P. K. Dash and D. S. Sisodia, "Transfer learning based lightweight model for classification of Alzheimer's disease using brain MR images," in *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0*, pp. 1–6, June 2024.

[124] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, June 2021.

[125] S. Vadera and S. Ameen, "Methods for Pruning Deep Neural Networks," *IEEE Access*, vol. 10, pp. 63280–63300, 2022.

[126] T. Hoefler, D. Alistarh, T. Ben-Nun, and N. Dryden, "Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks,"

[127] Y. Li, J. Luo, and J. Zhang, "Classification of Alzheimer's disease in MRI images using knowledge distillation framework: An investigation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, pp. 1235–1243, July 2022.

[128] M. Li, C. Cui, Q. Liu, R. Deng, T. Yao, M. Lionts, and Y. Huo, "Dataset Distillation in Medical Imaging: A Feasibility Study," Feb. 2025.

[129] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, and O. Colliot, "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," *Medical Image Analysis*, vol. 63, p. 101694, July 2020.

[130] V. M. Young, S. Gates, L. Y. Garcia, and A. Salardini, "Data Leakage in Deep Learning for Alzheimer's Disease Diagnosis: A Scoping Review of Methodological Rigor and Performance Inflation," *Diagnostics*, vol. 15, p. 2348, Sept. 2025.

[131] B. Tong, Z. Zhou, D. A. Tarzanagh, B. Hou, A. J. Saykin, J. Moore, M. Ritchie, and L. Shen, "Class-Balanced Deep Learning with Adaptive Vector Scaling Loss for Dementia Stage Detection," *Machine learning in medical imaging. MLMI (Workshop)*, vol. 14349, pp. 144–154, 2024.

[132] V. S. Diogo, H. A. Ferreira, D. Prata, and Alzheimer's Disease Neuroimaging Initiative, "Early diagnosis of Alzheimer's disease using machine learning: A multidiagnostic, generalizable approach," *Alzheimer's Research & Therapy*, vol. 14, p. 107, Aug. 2022.

[133] R. Turrisi, A. Verri, and A. Barla, "Deep learning-based Alzheimer's disease detection: Reproducibility and the effect of modeling choices," *Frontiers in Computational Neuroscience*, vol. 18, p. 1360095, Sept. 2024.

[134] A. L. Young, R. V. Marinescu, N. P. Oxtoby, M. Bocchetta, K. Yong, N. C. Firth, D. M. Cash, D. L. Thomas, K. M. Dick, J. Cardoso, J. van Swieten, B. Borroni, D. Galimberti, M. Masellis, M. C. Tartaglia, J. B. Rowe, C. Graff, F. Tagliavini, G. B. Frisoni, R. Laforce, E. Finger, A. de Mendonça, S. Sorbi, J. D. Warren, S. Crutch, N. C. Fox, S. Ourselin, J. M. Schott, J. D. Rohrer, and D. C. Alexander, "Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference," *Nature Communications*, vol. 9, p. 4273, Oct. 2018.

[135] A. Kumar, J. Sidhu, F. Lui, and J. W. Tsao, "Alzheimer Disease," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025.

[136] R. Turrisi, A. Verri, and A. Barla, "Deep learning-based Alzheimer's disease detection: Reproducibility and the effect of modeling choices," *Frontiers in Computational Neuroscience*, vol. 18, p. 1360095, Sept. 2024.

[137] M. J. Leming, E. E. Bron, R. Bruffaerts, Y. Ou, J. E. Iglesias, R. L. Gollub,

and H. Im, "Challenges of implementing computer-aided diagnostic models for neuroimages in a clinical setting," *NPJ Digital Medicine*, vol. 6, p. 129, July 2023.

[138] M. L. Raza, S. T. Hassan, S. Jamil, N. Hyder, K. Batool, S. Walji, and M. K. Abbas, "Advancements in deep learning for early diagnosis of Alzheimer's disease using multimodal neuroimaging: Challenges and future directions," *Frontiers in Neuroinformatics*, vol. 19, May 2025.

[139] S. Golriz Khatami, C. Robinson, C. Birkenbihl, D. Domingo-Fernández, C. T. Hoyt, and M. Hofmann-Apitius, "Challenges of Integrative Disease Modeling in Alzheimer's Disease," *Frontiers in Molecular Biosciences*, vol. 6, p. 158, Jan. 2020.

Abbreviations and Acronyms

AD	Alzheimer's Disease
AI	Artificial Intelligence
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CT	Computed Tomography
DL	Deep Learning
fMRI	Functional Magnetic Resonance Imaging
Grad-CAM	Gradient-weighted Class Activation Mapping
MCI	Mild Cognitive Impairment
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NC	Normal Control
PET	Positron Emission Tomography
ROI	Region of Interest
VBM	Voxel-Based Morphometry