SC42140 Signal Analysis & Learning for Two-Dimensional Systems

Practical Assignment Spring 2022

Version 1.0 – February 10, 2022

L. E. M. Tideman

Assignment deliverables: There are three parts to this practical assignment: "Part I: Clinical metadata analysis", "Part II: Exploratory analysis of MRI data", and "Part III: Convolutional neural networks". To obtain a grade for this assignment, you are required to submit a written report that includes your answers to the questions posed in Part I and Part II of the assignment (see below). Your report should be saved in PDF-format, and the filename should include your surname and your TU Delft student number. For Part III, you are required to submit your code to us (in ipynb-format). You are also required to upload your convolutional neural network (CNN) model's predictions to Kaggle (in csv-format), where your model's performance will be graded automatically. Summarizing, there are three deliverables: a written report answering questions specified below (Part I+II), your code for building your CNN-model for recognizing Alzheimer's disease (Part III), and your model's predictions should be uploaded to Kaggle to assess the performance you achieved (Part III).

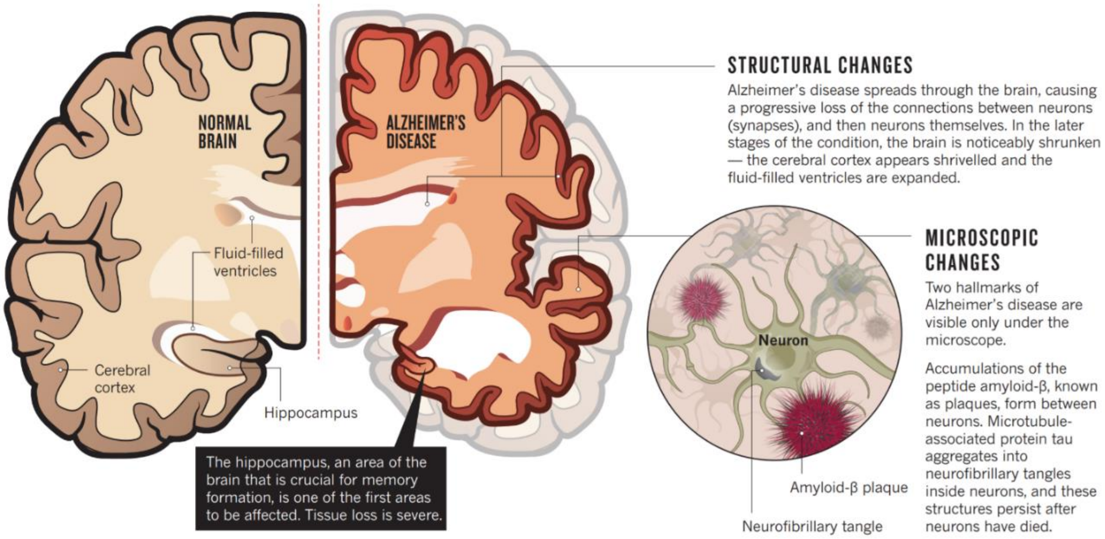
Submission deadline: 17h on the 29th of April 2022.

Table of contents

Introduction

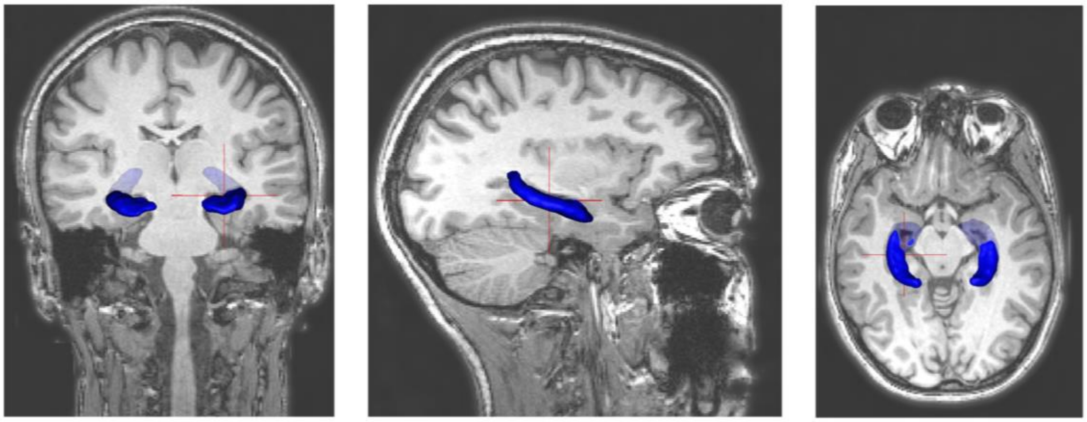
Alzheimer's disease

Alzheimer's disease (AD) is a neurodegenerative disease that is characterized by a progressive deterioration in cognitive function. AD affects approximately 50 million people worldwide and is expected to affect 1 out of 85 people globally by 2050 [Brookmeyer, 2007]. It is the most common form of dementia: it accounts for 60-70% of worldwide cases [WHO]. The causes of AD are not well understood, and no cure exists (yet). AD is one of the major causes of disability and dependency among elderly people, and adequately caring for AD patients is expected to increase the burden for national healthcare systems. In the United Kingdom for example, AD affects 1 in every 6 people over the age of 80 [NHS]. The prevalence of Alzheimer’s disease in people aged 60 or above is highest in North Africa and the Middle East [Drew, 2018]. Women are more likely to develop AD than men: in the United States, 2/3rd of people living with AD are women [ALZ]. The rate at which AD symptoms progress from mild to severe cognitive impairment differ from one person to another. AD is associated with neuronal loss in the medial temporal lobe and hippocampus [Gerardin, 2012; Salvatore, 2015]. Pathological changes usually begin and are ultimately most severe in the hippocampus, which is a brain structure associated with regulating emotions and forming new memories.


NORMAL BRAIN
ALZHEIMER'S DISEASE
Fluid-filled ventricles
Cerebral cortex
Hippocampus
STRUCTURAL CHANGES
Alzheimer's disease spreads through the brain, causing a progressive loss of the connections between neurons (synapses), and then neurons themselves. In the later stages of the condition, the brain is noticeably shrunken — the cerebral cortex appears shrivelled and the fluid-filled ventricles are expanded.
The hippocampus, an area of the brain that is crucial for memory formation, is one of the first areas to be affected. Tissue loss is severe.
Neuron
MICROSCOPIC CHANGES
Two hallmarks of Alzheimer's disease are visible only under the microscope.
Accumulations of the peptide amyloid-β, known as plaques, form between neurons. Microtubule-associated protein tau aggregates into neurofibrillary tangles inside neurons, and these structures persist after neurons have died.
Amyloid-β plaque
Neurofibrillary tangle

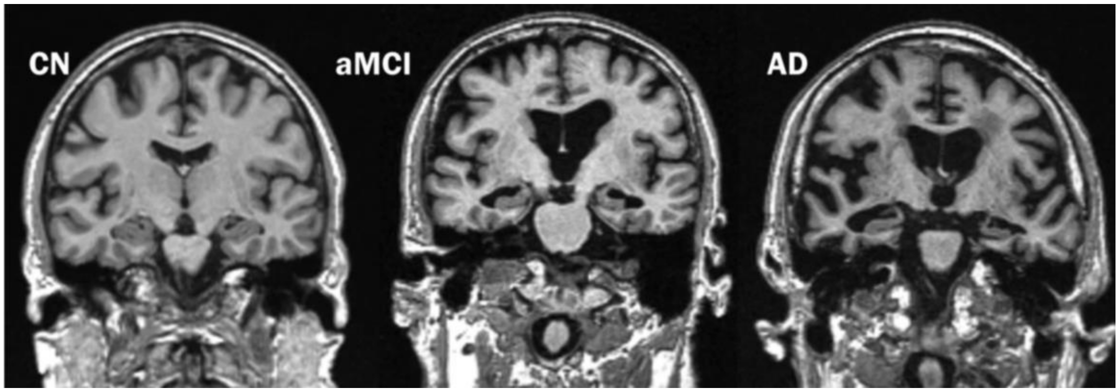
Healthy brain versus Alzheimer’s disease brain [Drew, 2018]. Atrophy is especially severe in the hippocampus.

Neuroimaging is increasingly used for the early diagnosis of AD. Magnetic resonance imaging (MRI) offers the possibility to study pathological brain changes associated with AD *in vivo*. MRI is a non-invasive medical imaging technique that uses strong magnetic fields and radio waves to produce three-dimensional (3D) anatomical images [MayoClinic]. A 3D MRI image is a 3D array of voxels, or volumetric pixels, whose size determines the image’s spatial resolution. In this assignment, you will study T1-weighted MRI images of the brain (as opposed to T2-weighted images). Because T1-weighted MRI highlights fatty tissue, white matter (myelinated axons) appears brighter than gray matter (neurons) [UCSanDiego; StanfordMRI].



Coronal, sagittal, and axial MRI views of the hippocampus [Gerardin, 2012].

In the context of AD, T1-weighted MRI is useful to visualize and measure cerebral atrophy (shrinkage due to the death of brain cells) in the hippocampus. For example, MRI-based volumetry of the gray matter in the hippocampal region has proven to be a reliable marker of moderate to severe AD [Gerardin, 2012; Salvatore, 2015]. Supervised deep learning methods, especially convolutional neural networks (CNNs), can be used to detect AD by automating the extraction of low-to-high level latent features from MRI data [Wen, 2020].



MRI scans showing the progressive atrophy of the medial temporal lobes in an older cognitively normal (CN) subject, an amnestic mild cognitive impairment (aMCI) subject, and an Alzheimer's disease (AD) subject [Vemuri, 2010].

Open-source MRI data

The Open Access Series of Imaging Studies (OASIS) is a publicly available series of neuroimaging datasets that are used for the study of Alzheimer's disease. In this assignment, you will be studying anatomical MRI data from the OASIS–1 cohort: it is a cross-sectional collection of 3D T1-weighted MRI scans from 377 right-handed subjects aged 18 to 96 [Marcus, 2007]. There is one 3D MRI image per subject and each subject has a patient ID. The diagnoses and clinical metadata of the study participants are also provided. A subject's diagnosis may be either CN (cognitively normal) or AD (Alzheimer's disease). You will have the opportunity to work on both raw and processed MRI data. The data is available in the course Google Drive.

Kaggle

You will use Kaggle for Part III of the assignment. Kaggle is an online data science community that serves as a platform for machine learning competitions on public datasets. In Part III, you need to solve a binary classification problem and your submission is graded automatically in Kaggle. In the context of SC42140, we have a non-public community competition entitled *SC42140 Part III - TU Delft 2022* that you can only access via an invitation link. Submitting a prediction requires uploading a comma-separated values text file, or csv-file, to Kaggle. The Kaggle competition's leaderboard will show you the classification score that you obtain on the testing set, as well as the scores of other students. In the context of SC42140, the public and private leaderboards are identical. You are allowed to make up to 20 submission per day. If you click on the My Submissions tab you will see a list of all your submissions. The metric used to evaluate the predictive performance of your classification model is the F1-score, which is the harmonic mean of precision and recall (with equal weighting of precision and recall). Your final score on performance at the end of the competition (29th of April) will be determined by whichever of your submissions performed best on the leaderboard. So, feel free to submit intermediately, to keep improving, and to submit better performing predictions later on (as long as the competition has not closed yet).

Google Colaboratory

Google Colaboratory, or Colab for short, is a hosted Jupyter notebook service from Google Research that provides you with free access to computing resources and does not require any software installations. You can enter Colab by double-clicking one of the ipynb-files in the shared Google Drive directory. The resources you have access to include a central processing unit (CPU), a graphics processing unit (GPU), and memory (up to 12GB of RAM). Jupyter notebooks allow you to combine text (Markdown) and Python code. Some remarks about Colab:

- The computing resources available via Colab are not guaranteed. The amount of memory available to you in Colab virtual machines varies over time and depends on the users whom you are sharing resources with. After connecting to a runtime (top right of your notebook), you can see the RAM and Disk resources available to you.

- Colab will disconnect your notebook, causing you to lose your local files (but not your code), if you leave it idle for more than 30 minutes. Your notebook will be automatically saved to your Google Drive.
- Using a GPU is not necessary for Part I and Part II of the assignment. However, it is recommended that you use a GPU for Part III. Make sure that you change the runtime type before starting Part III. Click on the Runtime tab, select Change Runtime Type in the drop-down menu, then set your hardware accelerator to a GPU.
- Since Colab notebooks are hosted on Google's cloud servers, there's no direct access to files on your local drive. We suggest that you download the data necessary for the assignment into your Google Drive, and then mount your Google Drive in the runtime's virtual machine.
- If you do not have/want a Google account, or if you prefer working on your own laptop rather than on the cloud, you can download the Colab notebooks locally as well. We suggest that you save them as a Jupyter notebook (.ipynb file format). Make sure to install the necessary Python libraries on your laptop, for example using Anaconda and PIP.

In the Google Drive, we have provided notebooks for each of the three parts of the assignment, with basic code to accompany the questions posed below and to get you started with writing your own code and building your model. Please follow along in your notebooks as you read through the questions below.

Questions

In your report, please label each answer with the number of the question it is answering.

Part I: Clinical metadata analysis

You will find the patients' clinical metadata in OASIS_data.csv in the Google Drive. Your task in Part I is to solve a binary classification problem whose aim is to differentiate the cognitively normal patients from those suffering from Alzheimer's disease. The metadata provides information about each patient's demographics (age and sex) and each patient's cognitive scores. Clinicians can use cognitive scores, such as the Mini-Mental State Evaluation (MMSE) and the Clinical Dementia Rating (CDR), to track disease progression. Note that, although assessment protocols are standardized, these scores can vary depending on the examiner and on the physical and/or mental condition of the patient during the test. The MMSE score ranges from 0 (severe dementia) to 30 (no dementia), whereas the CDR ranges from 0 (no dementia) to 3 (severe dementia). You therefore have three numeric features (age, MMSE score, CDR score) and one categorical feature (sex) on the basis of which to determine if a given patient has Alzheimer's disease.

1. Use the Pandas data analysis library to explore the clinical data. The clinical data is provided to you in the form of a Pandas DataFrame: it is a two-dimensional tabular data structure with rows (one row per patient and MRI dataset) and columns (one column per feature). You can include the last column, which corresponds to the patients' CN or AD diagnosis. Use the info method to obtain a summary of the clinical data. Use the describe method to obtain statistics about the central tendency, dispersion, and shape of a dataset's distribution. Answer the following questions about the patient cohort and dataset:
   1.1. A missing value is denoted by NaN (Not a Number). You can use the isnull method to detect missing values. Which feature in the dataset has missing values? The data is not missing at random. Given the distribution of missing values, how do you suggest imputing the missing values?
   1.2. A dataset is balanced if the number of patients in each class (cognitively normal and Alzheimer's disease) is similar. Is the dataset balanced or imbalanced? Based on the cognitive scores of diseased patients, estimate how many patients have severe Alzheimer's disease.
   1.3. What is the overall age distribution of patients? What percentage of the MRI data corresponds to female patients? What is the median age of healthy and diseased patients?
   1.4. Use the Matplotlib and Seaborn data visualization libraries to visualize the data. Include one histogram and one other figure (e.g. scatterplot, violin plot) in your answers to the previous questions. Each figure should have a title, and its axes must be labelled.

2. Use unsupervised machine learning methods from the scikit-learn library to study the data. Make sure that you exclude the diagnosis information (CN or AD). Use the k-means clustering

algorithm (or another algorithm of your choice) to identify several groups of similar patients. Try out different combinations of the hyper-parameters (number of clusters, distance metric).

2.1. Explain the clustering algorithm that you use. How is the clustering initialized? How is it updated? What measure is used to quantify clustering quality?

2.2. Describe the results of your analysis. Use Matplotlib and Seaborn to visualize the results of your analysis. Which feature drives the clustering? Do you find the results of clustering useful? Do you observe any outliers (abnormal patients)?

3. Follow the instructions in the notebook to train, validate, and test a simple classification model using the scikit-learn library. You will be using a logistic regression model for classification. A logistic regression model is a generalized linear model (GLM) with a logit linking function and a binary response. You are not required to write any code in question I.3 (this question). The purpose of question I.3 is for you to understand the different steps of supervised machine learning: data processing (scaling of features in our case), model architecture choice (logistic regression in our case), model parameter optimization (also called training), and, finally, model predictive performance assessment (also called testing).

3.1. Explain how a logistic regression model is obtained from labelled data, and how it performs classification of new unlabelled data. Include the formula of how the probability of one patient having Alzheimer's disease is a linear model of the log odds.

3.2. Study the coefficients of the logistic regression model. There are four coefficients corresponding to the four features (sex, age, CDR score, and MMS score). How do you interpret these coefficients? *Hint: think about positive and negative correlations.*

3.3. Look at the model accuracy and confusion matrix. Why is the model performance so good? Given the data we dispose of, would such a classification task be useful in a clinical setting? *Hint: look into data leakage.*

Part II: Exploratory analysis of MRI data

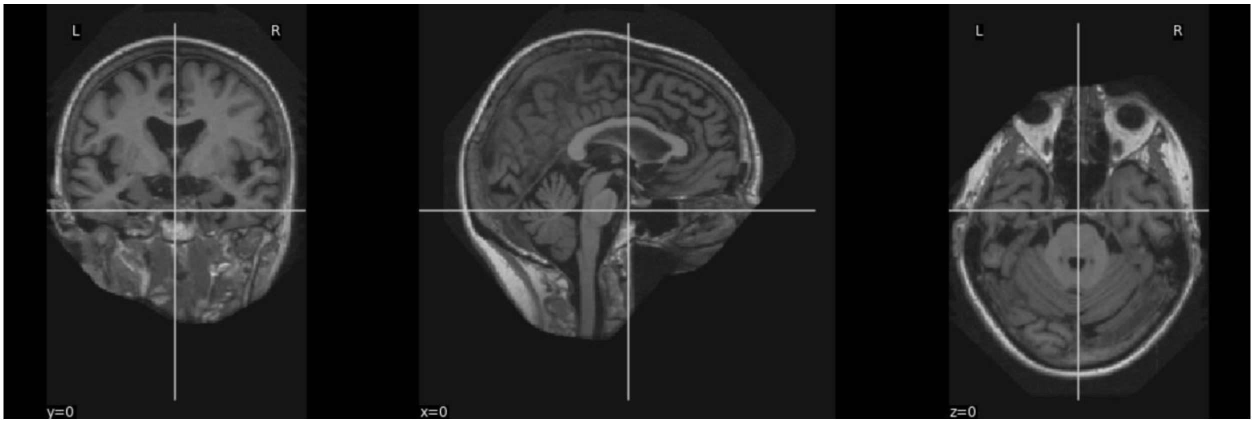
Each patient's raw 3D MRI data is provided to you in a compressed NIfTI-1 neuroimaging file format with a ".nii.gz" file extension. NIfTI is short for Neuroimaging Informatics Technology Initiative, and it is a commonly used format for neuroimaging. In the course Google Drive, you are provided with the MRI scans of five patients: OASIS10001, OASIS10003, OASIS10184, OASIS10322, and OASIS10440. These MRI scans respectively correspond to patient numbers 281, 243, 231, 221, and 198 in the clinical data (OASIS_data.csv). The following questions relate to the MRI scan of patient OASIS10003. You can experiment with the other datasets if you like.

1. We use a Python library called NiBabel to load NIfTI images into a NumPy array. A NiBabel image object is the association of an image data array, an affine array with the position of the image array data in a reference space, and image metadata in the form of a header. The

reference space is defined relative to the magnet isocentre of the scanner: the magnet isocentre is the origin and the three axes are respectively called the scanner-bore axis (X), the scanner-floor/ceiling axis (Y) and the scanner-left/right (Z). Check out the NiBabel documentation and answer the following questions about the image header:

1.1. What are the dimensions of the image data?
1.2. What size are the voxels (specify the real-world unit)?
1.3. What is the data type of the image data?

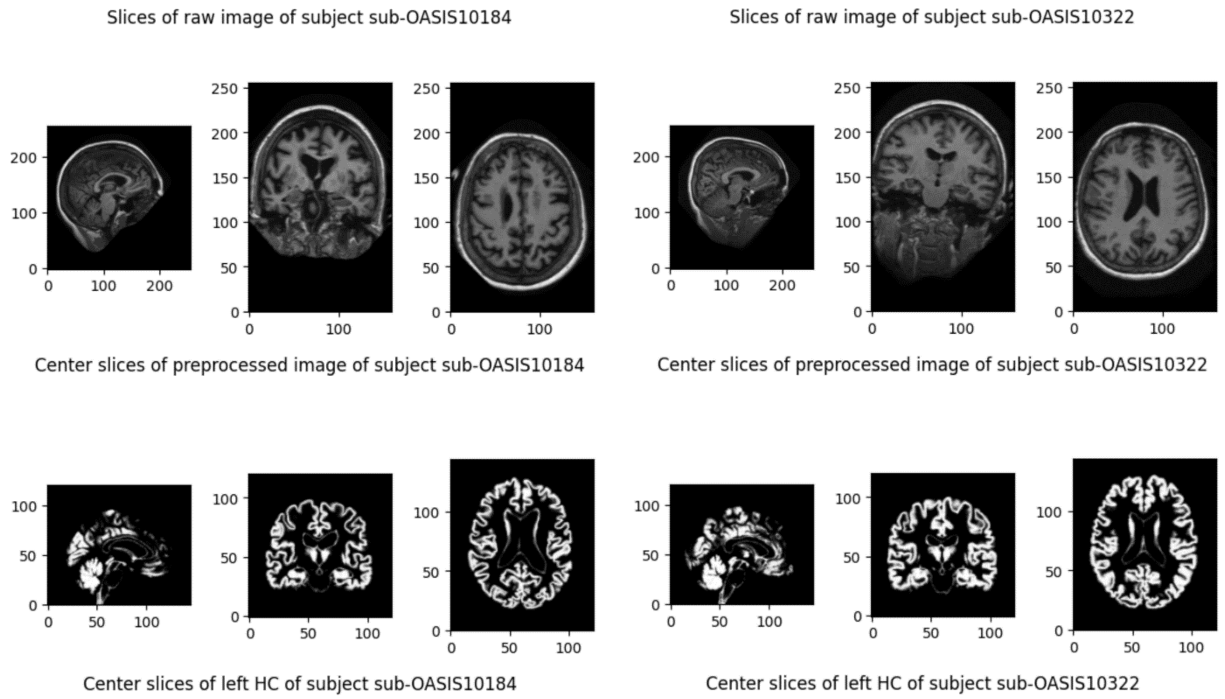
2. Inspect the MRI data by slicing the 3D tensor through the middle of the three different spatial axes. Use the Matplotlib library and follow the instructions in the notebook to slice the 3D MRI dataset and visualize the MRI signal intensity.
   2.1. Which of the three X, Y, Z axes corresponds to the sagittal, axial, coronal brain sections?
   2.2. For each of the five patient datasets, put the three brain slices that you obtain in your report. Your figures should resemble the figure hereunder.



3. Follow the instructions in the notebook to improve the background mask obtained by a simple intensity thresholding operation. You can use the multidimensional image processing functions (e.g. smoothing, filtering, or morphological operations) provided by Scipy.
   3.1. Briefly explain what image processing methods you use.
   3.2. Include a figure of the mask you obtain in your report.

Part III: Convolutional neural networks

Although deep learning allows for the analysis of raw data, the T1-weighted MRI data files studied in Part III are the result of extensive image processing. The processing was done using two open-source software packages for clinical neuroimaging studies: SPM by the University College London, and Clinica by the Aramis Lab and the Paris Brain Institute. The main processing steps are the following: conversion from NIfTI to BIDS format, spatial normalization, non-linear registration, and segmentation of grey matter, conversion from BIDS to tensor format, and finally cropping of the left hippocampus. For each patient (or subject) in the OASIS–1 cohort, you are provided with a 3D tensor of the patient's left hippocampus. The left hippocampus is a neuroanatomical region that is atrophied for most AD right-handed patients. All processed MRI datasets are smaller than the raw MRI datasets and have the same size (30x40x30). Limiting our scope to the left hippocampus is necessary given that deep learning is computationally expensive, and that Google Colaboratory provides you with limited resources.



Slices of raw image of subject sub-OASIS10184
Slices of raw image of subject sub-OASIS10322
Center slices of preprocessed image of subject sub-OASIS10184
Center slices of preprocessed image of subject sub-OASIS10322
Center slices of left HC of subject sub-OASIS10184
Center slices of left HC of subject sub-OASIS10322
0
50
100
150
200
250

Comparison of raw (top) and processed (bottom) MRI scans of two OASIS subjects

Part III does not require you to answer any questions. It does require you to study the PyTorch tutorial that is provided in the form of a Google Colaboratory notebook. You will also need to use the PyTorch documentation about convolution layers, pooling layers, non-linear activations, and loss functions. You will then be able to fill in the missing code in the Part III notebook (you can modify the whole notebook if you like). Your aim is to minimize the diagnosis errors made by the convolutional neural network that you design to classify the 3D MRI scans of patients' left hippocampus. Given the training data imbalance, you need a minimum accuracy of 80% to get a passing grade for Part III of the assignment. You can account for the data imbalance by weighing classes in the loss function or by having the data loader oversample the minority class. Predicting that a subject has Alzheimer's disease is called a positive prediction (True), whereas predicting that a subject is cognitively normal is called a negative prediction (False). Further instructions are provided in the SC42140 community Kaggle competition. The training dataset (220 subjects) and testing dataset (157 subjects) are also available in the Google Drive and via Kaggle (see the Data tab).

References

[Brookmeyer, 2007] "Forecasting the global burden of Alzheimer's disease" by Ron Brookmeyer, Elizabeth Johnson, Kathryn Ziegler-Graham, and H. Michael Arrighi. *Alzheimer's & dementia: the journal of the Alzheimer's association.* Volume 3, July 2007. https://doi.org/10.1016/j.jalz.2007.04.381

[Drew, 2018] "An age-old story of dementia" by Liam Drew, *Nature*. Volume 559, July 2018. https://www.nature.com/articles/d41586-018-05718-5

[WHO] Dementia fact sheet – World Health Organization https://www.who.int/en/news-room/fact-sheets/detail/dementia

[NHS] Alzheimer's disease overview – National Health Service, UK https://www.nhs.uk/conditions/alzheimers-disease/

[ALZ] Women and Alzheimer's – Alzheimer's association https://www.alz.org/alzheimers-dementia/what-is-alzheimers/women-and-alzheimer-s

[MayoClinic] Magnetic resonance imaging – Mayo Clinic https://www.mayoclinic.org/tests-procedures/mri/about/pac-20384768

[UCSanDiego] Structural MRI Imaging – UC San Diego, School of Medicine, Department of Radiology http://fmri.ucsd.edu/Howto/3T/structure.html

[StanfordMRI] Magnetic Resonance Imaging by the Neuro-ophthalmology School of Medicine, Stanford University https://neuro-ophthalmology.stanford.edu/2017/05/neuro-ophthalmology-question-of-the-week-magnetic-resonance-imaging/

[Gerardin, 2012] "Morphometry of the human hippocampus from MRI and conventional MRI high field" by Emilie Gerardin. PhD thesis. Université Paris Sud – Paris XI, September 2013.

https://tel.archives-ouvertes.fr/file/index/docid/856589/filename/VD2_GERARDIN_EMILIE_13122012.pdf

[Vemuri, 2010] “Role of structural MRI in Alzheimer's disease” by Prashanthi Vemuri, and Clifford R. Jack Jr. *Alzheimer's Research & Therapy*, Volume 2, August 2010. https://doi.org/10.1186/alzrt47

[Markus, 2007] “Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults” by Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, Randy L. Buckner. *Journal of Cognitive Neuroscience*, Volume 19, September 2007; https://doi.org/10.1162/jocn.2007.19.9.1498

[Salvatore, 2015] “Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach” by Christian Salvatore, Antonio Cerasa, Petronilla Battista, Maria C. Gilardi, Aldo Quattrone, Isabella Castiglioni, and Alzheimer's Disease Neuroimaging Initiative. *Frontiers in Neuroscience.* Volume 9, September 2015. https://doi.org/10.3389/fnins.2015.00307

[Routier, 2021] “Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies” by Alexandre Routier, Ninon Burgos, Mauricio Díaz, Michael Bacci, Simona Bottani, Omar El-Rifai, Arnaud Marcoux, Tristan Moreau, Jorge Samper-González, Marc Teichmann, Elina Thibeau-Sutre, Ghislain Vaillant, Junhao Wen, Adam Wild, Marie-Odile Habert, Stanley Durrleman, and Olivier Colliot. *Frontiers in Neuroinformatics*; Volume 15, August 2021. https://doi.org/10.3389/fninf.2021.689675

[Wen, 2020] “Convolutional Neural Networks for Classification of Alzheimer’s Disease: Overview and Reproducible Evaluation” by Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, and Olivier Colliot. *Medical Image Analysis*, Volume 63, July 2020. https://doi.org/10.1016/j.media.2020.101694

[LeCun, 2015] “Deep learning” by Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. *Nature* Vol 521, May 2015. https://www.nature.com/articles/nature14539

Additional deep learning resources

- *Deep Learning* – online book by Ian Goodfellow, Yoshua Bengio and Aaron Courville https://www.deeplearningbook.org/
- *Neural Networks and Deep Learning* – online book by Michael Nielsen http://neuralnetworksanddeeplearning.com/
- "Deep learning" by Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. *Nature* Vol 521, 436–444 (May 2015). https://www.nature.com/articles/nature14539
- Introduction to Deep Learning (MIT 6.S191, Massachusetts Institute of Technology, Spring 2021) – online course by Alexander Amini and Ava Soleimany http://introtodeeplearning.com/ & https://www.youtube.com/watch?v=5tvmMX8r_OM&ab_channel=AlexanderAmini & https://www.youtube.com/watch?v=AjtX1N_VT9E&ab_channel=AlexanderAmini
- Deep Learning Essentials – online edX course by IVADO, Mila and the University of Montreal https://www.edx.org/course/deep-learning-essentials
- Convolutional Neural Networks for Visual Recognition (Stanford CS231n, Stanford University, Spring 2021) https://cs231n.github.io/
- Introduction to Deep Learning in PyTorch – Lecture by Evann Courdier (EPFL) https://dl4sci-school.lbl.gov/evann-courdier & https://github.com/theevann/dl4sci-pytorch-webinar
- Deep Learning (NYU DS-GA 1008, New York University, Spring 2020) – Course by Yann LeCun and Alfredo Canziani https://atcold.github.io/pytorch-Deep-Learning/ & https://www.youtube.com/playlist?list=PLLHTzKZzVU9eaEyErdV26ikyolxOsz6mq
- Deep Learning with PyTorch Step-by-Step: A Beginner's Guide – online book by Daniel Voigt Godoy https://github.com/dvgodoy/PyTorchStepByStep