

École polytechnique de Louvain

Machine learning for early prediction of the onset of Alzheimer's dementia using MRI, PET and cognitive tests

Author: **Mathias MARQUES**

Supervisor: **Benoît MACQ**

Readers: **Bernard HANSEEUW, Laurence DRICOT, Colin VANDEN BULCKE**

Academic year 2022–2023

Master [120] in Computer Science and Engineering

Abstract

Context: Alzheimer's disease's rising incidence poses a significant challenge for global healthcare systems, progressively impairing cognitive abilities and leading to dementia. Its prevalence increases with age, exacerbating concerns as global life expectancy rises.

Objective: The thesis aims to develop a predictive model determining the onset of dementia in Alzheimer's disease's patients accurately. In addition, a secondary goal is to distinguish between patients who will progress to dementia and those who won't.

Methods: The research starts by meticulously analysing the relevant data sources, including MRI, PET scans and cognitive test data. Initially, the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm and Principal Component Analysis (PCA) will be used to visualise the data. Following this, the study will delve into diverse machine learning classifiers like Support Vector Machines (SVM) and Multilayer Perceptron (MLP) models.

Results: The best model found used an MLP and achieved an accuracy of 0.6667 for predicting the onset of dementia. When trying to predict whether the patient will progress to dementia, an SVM achieved the highest accuracy with 0.9524.

Conclusion: The models developed in this thesis have limitations due to data class imbalance, which affects the accuracy of dementia onset prediction. Addressing this imbalance is crucial to improve the accuracy of dementia onset prediction. However, the model for predicting whether a patient will progress to dementia is very accurate and shows promising results in this specific area.

Acknowledgements

During my dissertation I received a lot of help. This small section is here to thank the people who supported me.

I would like to thank my supervisor, Benoit Macq for his time and valuable insights on the research and the framework it has put in place to help me achieve my objectives.

I would like to thank the research team of the Woluwe campus for introducing me to the medical field. Particularly Bernard Hanseeuw for the help on the interpretation on some of the result. I am also grateful to Laurence Dricot, Lise Colmant, Vincent Malotaux, Yasmine Salman, Lisa Quenon, Lara Huyghe, Colin Van Eycken, and Emilien Boyer for their invaluable guidance in comprehending the collected data across various studies.

I wish to thank particularly Colin Vanden Bulcke for his help throughout this thesis and especially for his invaluable help and advice during the writing process.

I would like to thank my family, my girlfriend and my friends for their unfailing support throughout this Masters. This achievement wouldn't have been possible without their encouragement and assistance. Specifically, I extend my thanks to Daniel Marques, Marie Sion and Julien Herman for reviewing my dissertation.

To all of you thank you.

List of Figures

1	Pathophysiology of AD	5
2	Position of the entorhinal cortex and the hippocampus in the brain	6
3	Desikan-killiany and Destrieux atlas	9
4	PCA projection	13
5	Separation of the data for machine learning	15
6	Kfold	16
7	Illustration of the SVM features	16
8	Neural network of an MLP	18
9	Perceptron operation	18
10	Demographic characteristics of the Biomarker study	22
11	Demographic characteristics of the TAU study	23
12	Pipeline from an MRI image to the csv files	26
13	Pipeline from a PET scan image to the csv files	27
14	Split of the cognition tests data	28
15	Database Architecture	29
16	Explained variance of the PCA component	33
17	Architecture of the MLP used in this thesis	35
18	t-SNE in 2 dimensions with the data from the MRI and both PET with the modified labels	37
19	t-SNE in 2 dimensions with the data from the MRI and both PET with the original labels	37
20	t-SNE in 3 dimensions with the data from the MRI and both PET scan with the original labels	38
21	t-SNE in 2 dimensions on all the data with the modified labels . . .	38
22	PCA in 2 dimensions with the data from the MRI and both PET with the modified labels	39
23	PCA in 2 dimensions with the data from the MRI and both PET with the original labels	39
24	PCA in 3 dimensions with the data from the MRI and both PET scan with the modified labels	40

25	PCA in 3 dimensions with the data from the MRI and both PET scan with the original labels	40
26	PCA in 2 dimensions on all the data with the modified labels	41
27	ROC curves of the binary classification made by a SVM. The binary classification aims to predict whether patients will develop dementia specifically due to AD	42
28	Combined confusion matrix of 10 runs to predict the onset of dementia. Prediction made by an SVM with the z-score normalization and all the data.	43
29	Combined confusion matrix of 10 runs to predict the onset of dementia. Prediction made by an SVM with the min max normalization and all the data.	43
30	Combined confusion matrix of 10 runs to predict the onset of dementia. Prediction made by an SVM with the z-score normalization and all the data except the cognition test	44
31	Precision recall graph of one run to predict the onset of dementia with an SVM the z-score normalization and all the data	44
32	Combined confusion matrix of 10 runs to predict the onset of dementia. Prediction made by an SVM with the z-score normalization and all the data without cognition and without PCA.	45
33	Combined confusion matrix of 10 runs to predict the onset of dementia. Prediction made by an SVM with the z-score normalization and all the data and without PCA.	45
34	Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data without cognition and with PCA.	46
35	Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data with a PCA.	46
36	Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data without cognition and without PCA.	47
37	Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data without a PCA	47
38	Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data without cognition and with PCA	48

39	Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data with PCA	48
40	Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data without cognition and without PCA	49
41	Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data without a PCA	49
42	Illustration of a Freesurfer text file of a patient's left hemisphere (lh.aparc.DKTatlas)	70
43	Illustration of part of a Freesurfer text file for a patient's amyloid PET scan	72

List of Abbreviations

AD:	Alzheimer’s disease
MRI:	Magnetic Resonance Imaging
PET scans:	Positron Emission Tomography Scans
t-SNE:	t-Distributed Stochastic Neighbor Embedding
PCA:	Principal Component Analysis
SVM:	Support Vector Machines
MLP:	Multilayer Perceptron
APP:	Amyloid Precursor Protein
PSEN1:	Presenilin 1
PSEN2:	Presenilin 2
APOE:	Apolipoprotein E
MCI:	Mild Cognitive Impairment
MMSE:	Mini-Mental State Examination
TICS:	Telephone Interview for Cognitive Status
FCSRT:	Free and Cued Selective Reminding Test
TMT:	Trail Making Test
CERAD:	Consortium to Establish a Registry for Alzheimer’s Disease
ECog:	Everyday Cognition
DTI-MRI:	Diffusion MRI
FDG:	Fluorodeoxyglucose
SUVRs:	Standardized Uptake Value Ratios
PVC:	Partial Volume Correction
PVE:	Partial Volume Effects’
CSF:	Cerebrospinal Fluid
ROC:	Receiver Operating Characteristic
AUC:	Area Under the Curve
ADNI:	Alzheimer’s disease neuroimaging initiative
e.g.:	exempli gratia (for example)

Contents

Introduction	1
State of the art	2
1 Alzheimer's disease	3
1.1 Introduction	3
1.2 Epidemiology	3
1.3 Etiology	4
1.4 Pathophysiology	5
1.5 Diagnosis	6
1.6 Treatments	11
2 Algorithms	12
2.1 Visualization and dimension reduction Algorithms	12
2.2 Predictive Machine Learning Algorithms	15
3 Existing research	20
Methods and Materials	20
4 Materials	21
4.1 Presentation of the studies	21
4.2 MRI	25
4.3 PET scan	26
4.4 Cognition test	27
4.5 Database	29
5 Methods	30
5.1 Inclusion criteria	30
5.2 Preprocessing of the Data	31
5.3 Visualisation of the data	32
5.4 Machine learning models	32

Results	35
6 Visualization	36
6.1 t-SNE	36
6.2 PCA	39
7 Predictive Machine Learning Algorithms	41
7.1 SVM	41
7.2 MLP	45
Discussion	49
8 Visualization	50
9 Predictive Machine Learning Algorithms	52
9.1 SVM	52
9.2 MLP	53
10 limitation of the study	55
11 Perspectives	55
Conclusion	56
Appendix	66
12 Dataset creation	67
12.1 MRI Processing	67
12.2 PET scan Processing	71

Introduction

Alzheimer’s disease (AD) represents a major challenge for modern healthcare and places a heavy cost on individuals, families and healthcare systems around the world. This neurodegenerative disease progressively erodes cognitive functions, leading to a decline in memory, language, spatial perception and problem-solving abilities. Eventually, it culminates in a stage known as dementia, which significantly impairs day-to-day activities [1, 2].

The main cause behind dementia is AD, a condition whose likelihood of occurrence increases significantly with age. This is a growing concern as life expectancy increases around the world. This demographic shift reinforces the urgent need to understand, predict and treat this omnipresent disease, which is placing a heavy burden on society as a whole [3, 4].

While AD’s etiology encompasses various factors such as advancing age, genetic predispositions, lifestyle choices, and environmental influences, the central focus of this thesis lies in predicting the onset of dementia resulting from AD. Using machine learning and predictive models, this research aims to contribute to early intervention strategies and personalized care for individuals at risk of Alzheimer’s-related dementia. For example, if the model proves precise enough in predicting the disease’s progression, this method could enable a study on the potential impact of a medicine on an individual, paving the way for personalized therapeutic approaches and help to assess the effectiveness of a new drug.

As said above, the primary goal of this thesis is to create a model that can accurately predict the year in which a patient will advance to the stage of dementia due to AD. To accomplish these objectives, we will rely on various sources of data and explore several machine learning classifiers. Due to the data being available only at two-year intervals, it is not feasible to employ a regression model for the primary goal. Thus we will use classification methods instead of a regression techniques. Furthermore, a secondary objective is to differentiate between patients who will progress to the stage of dementia and those who will not.

To accomplish our objectives, we will initiate the process by carefully examining various data sources relevant to our research. These sources will include imaging data obtained from magnetic resonance imaging (MRI), positron emission tomography (PET) scans, as well as the data of the cognitive test. In the initial stage, we will use the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm and a Principal Component Analysis (PCA) to visualize and analyze the data. Subsequently, we will explore the implementation of various machine learning classifiers, including Support Vector Machines (SVM) and Multilayer Perceptron (MLP) models. By adopting this approach, our aim is to determine the most effective classification method for accurately predicting the onset of dementia.

The thesis is structured to comprehensively address the complexities of predicting the onset of dementia resulting from AD. The first chapter delves into the State of the Art, offering an extensive review of AD and existing predictive methodologies. This section also introduces the algorithms and predictive models employed in the subsequent analyses. Following this, the Material and Methods chapter details the methodologies employed in data collection, analysis, and model development. Subsequently, the Results chapter presents the findings and outcomes obtained from the predictive models. The Discussion chapter critically analyses the results and relates them to the literature. Lastly, the Conclusion encapsulates the study's key takeaways, emphasizing its contributions to early intervention strategies and personalized care in Alzheimer's related dementia.

State of the art

In this thesis some parts have been reformulated using ChatGPT [5] and the new tool of DeepL [6].

1 Alzheimer's disease

As previously stated, the main goal of this thesis is to forecast the onset of the critical stage of dementia resulting from AD. To achieve this, we will start by providing an overview and introduction to this medical condition.

1.1 Introduction

Understanding AD is one of the significant challenges in the medical world today. AD is a neurodegenerative condition that progresses over time, leading to a decline in both behavioral and cognitive functions. As the disease advances, it leads to a gradual deterioration in memory, language, spatial awareness, and problem-solving abilities [1]. After some time, the patient will be in a stage called dementia. This term is used to describe the condition where a patient's cognitive abilities are severely impaired, impacting their daily living activities [2]. AD correlates to the most common cause of dementia. With the increase in life expectancy worldwide, age-related cognitive disorders are becoming more prevalent, placing considerable burdens on individuals, families, and healthcare systems. This section will provide a comprehensive exploration of AD, covering its etiology, epidemiology, pathophysiology, symptoms, diagnosis, treatment, and consequences.

1.2 Epidemiology

As stated previously, AD is the most common form of dementia in the elderly, accounting for 60-80% of all dementia cases [3]. Its impact is far-reaching, affecting approximately 55 million people worldwide, and this number is projected to triple by 2050, constituting a significant global public health concern [4].

The increasing incidence of AD is closely linked to the aging population. As its prevalence continues to rise, understanding the disease becomes essential in enhancing the quality of life for those affected as well as guiding public health policies.

Moreover, the expenses associated with caring for individuals with dementia are substantial. In the United States, the annual payments for healthcare and long-term care for people with AD or other dementias are projected to skyrocket from \$345 billion in 2023 to nearly \$1 trillion in 2050, as reported by the Alzheimer's Association [7]. These escalating costs create economic burdens for families, governments, and healthcare systems worldwide, underscoring the urgency to focus on early diagnosis, combat AD, and seek effective treatments to prevent or slow its progression. By addressing these challenges, we can work towards alleviating the impact of Alzheimer's on individuals and society as a whole.

1.3 Etiology

There are multiple risk factors that increase the likelihood of developing AD. Age is considered to be one of the most important risk factors for cognitive decline and AD. It is estimated that up to 50% of individuals over 85 years old may be affected by the disease [8]. In addition to age, genetics also plays a significant role in the development of AD. Researchers have identified several genes that are associated with an increased risk of developing the disease. Including, the amyloid precursor protein (APP) gene, the presenilin 1 (PSEN1) gene, the presenilin 2 (PSEN2) gene, and the apolipoprotein E (APOE) gene [8]. Additionally, lifestyle (e.g. alcohol, smoking, ...) and environmental (e.g. air pollution, exposition to pesticide,... [9]) factors have been found to contribute to the development of the disease. Research has also shown that gender is a significant factor in the development of AD. According to a survey, nearly two-thirds of AD patients are female [10].

On the other hand, there are also factors that can reduce the risk of developing AD. Engaging in activities such as playing music, maintaining a healthy diet, participating in aerobic sports, and other lifestyle choices have been associated with a potential protective effect against the onset of AD [2].

Overall, AD is a complex disorder with multiple risk factors. It is important to note that having one or more of these risk factors does not guarantee that an individual will develop AD.

1.4 Pathophysiology

AD is a neurodegenerative disorder characterized by specific pathological changes in the brain. At a microscopic level, the main features of AD pathology are the presence of extracellular amyloid- β plaques and intracellular neurofibrillary tangles. Their accumulation leads to progressive neuronal loss and synaptic dysfunction, causing widespread brain atrophy and impaired neural communication [11]. In Figure 1, figure B illustrates the presence of extracellular amyloid- β and intracellular neurofibrillary tangles, alongside the depiction of brain atrophy in Figure C.

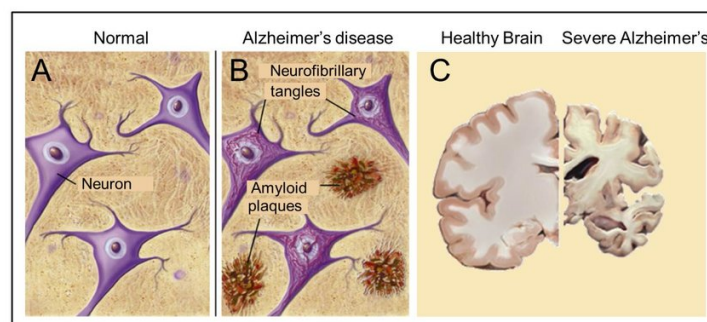


Figure 1: This image highlights the major features of the AD pathology: extra-cellular amyloid- β plaques and intracellular neurofibrillary tangles as well as the atrophy of the brain. Figure A : represents a healthy individual. Figure B : shows the amyloid- β plaques and the neurofibrillary tangles. Figure C : illustrate the brain atrophy caused by AD. **source A and B** are from google image **C** is from [12]. The combination of these figures adapted from [13].

Amyloid- β plaques refer to the abnormal buildup of proteins outside neurons in the brain, creating insoluble formations. These plaques are composed principally of aggregates of amyloid- β peptides and are considered a central feature of AD pathology. However, it's essential to note that they are not exclusive to AD, other diseases can also exhibit the development of amyloid- β plaques [14]. They disrupt normal cellular communication and contribute to neurotoxicity, leading to neuronal dysfunction and degeneration [2]. Amyloid deposition can initiate up to 20 years before the onset of the first symptoms of AD [15].

Neurofibrillary tangles, a pathological feature of AD, are abnormal structures found inside neurons in the brain. They are composed of hyperphosphorylated tau protein, which undergoes abnormal changes and aggregates, leading to destabilization of the neuronal cytoskeleton. The primary role of the tau protein is to stabilize axonal microtubules. Consequently, the presence of these tangles causes affected neurons to become dysfunctional and eventually leads to their death [2].

The exact mechanisms underlying the formation and propagation of amyloid- β plaques and neurofibrillary tangles in AD remain an area of intense and ongoing research [16] [17]. One of the current proposition is the amyloid cascade hypothesis. This assumption proposes that the accumulation of amyloid- β is a key event leading to AD by triggering tau hyperphosphorylation, neurofibrillary tangle formation, and neurodegeneration [18].

In AD, the first affected regions are the hippocampus and the entorhinal cortex. The anatomical positions of the entorhinal cortex (in red) and hippocampus (in blue) are displayed in the figure 2. These brain areas play critical roles in memory formation and consolidation. This explains that memory is particularly vulnerable to AD. As the disease progresses, the damage extends to other parts of the brain, leading to widespread cognitive decline and memory impairment [19].

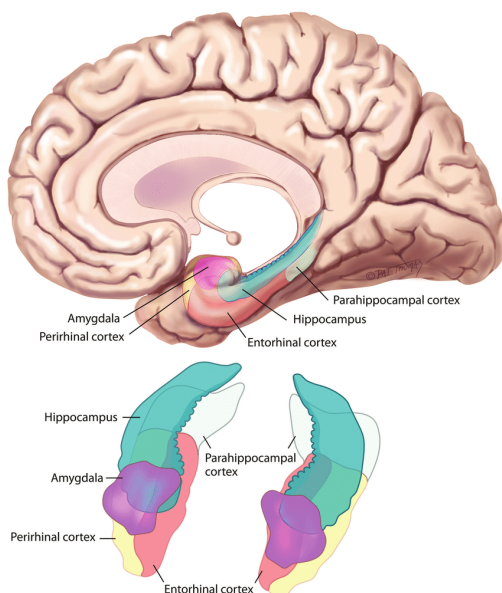


Figure 2: An illustration of the brain to indicate the position of the entorhinal cortex and the hippocampus. The combination of these figures adapted from [13] [20].

1.5 Diagnosis

Early diagnosis of AD is crucial, whether it's for medical testing or end-of-life planning, such as inheritance matters. Although a definitive diagnosis is only possible post-mortem, clinical diagnosis can be established through Neuropsychological tests or by Biomarkers measurements. The Neuropsychological test will be described later. Then, quantitative ways of diagnosis will be explored through

biomarkers. According to [21], A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. Numerous biomarkers have been identified for AD diagnosis. They will be discussed in the context of MRI, PET, and fluid analysis. Following this, the challenges associated with this diagnosis will be briefly discussed.

1.5.1 Neuropsychological Test

Neuropsychological tests are used to assess the cognitive impairment and classify its advancement into different stages: no cognitive impairment, subjective cognitive impairment, mild cognitive impairment (MCI), and the final stage dementia. To establish this classification the clinician can use a wide battery of tests. The tests used in this study are presented below. Please note that this list isn't exhaustive and doesn't cover all potential tests that are possible to evaluate the cognition of a patient.

- **Mini-Mental State Examination (MMSE):** A general assessment for cognitive impairment, widely used by psychologists [22].
- **Telephone Interview for Cognitive Status (TICS):** Allows remote testing for a general overview of cognitive state [23].
- **Free and Cued Selective Reminding Test (FCSRT):** A memory related test that has been proven to greatly distinguish AD from other memory related disease [24].
- **Luria Test:** Evaluates executive capabilities by reproducing hand movement sequences [25].
- **Trail Making Test (TMT):** Assesses visual perception, tracking speed, executive function, and task switching [26].
- **Naming Lexis Test:** Evaluates a patient's language by asking them to name presented objects [27].
- **Fluency Test:** Assesses language in two parts, asking for animal names and words starting with a specific letter [28].
- **Clock Test:** Involves drawing a clock displaying a specific time to assess cognitive function [29].

- **Consortium to Establish a Registry for Alzheimer’s Disease (CERAD):** Measures cognitive impairment for AD through various tests [30].
- **Everyday Cognition (ECog) Tests:** Evaluates cognitive abilities in real-world situations, providing insights into daily cognitive performance [31].
- **Logical memory test from the Wechsler Memory Scale:** In the first part of this test, a story is told to the patient and he has to immediately tell it back. In the second part after some time the patient has to tell the story again. Then some yes or no questions are asked at the end of the story. [32]
- **The shape test form Doors and People tests:** In this test, 4 simple shapes are presented to the patient in the first time it is asked to the patient to copy them. And in a second time he has to draw the shape from memory [33].

1.5.2 Magnetic Resonance Imaging

To diagnose AD, the physicians rely on the MRI and other imaging techniques as it will be described later. An MRI is a non-invasive imaging method that uses a powerful magnetic field to align protons within the body. Subsequently, it emits radio-frequency pulses, disrupting the alignment established by the magnet. A sensor then captures the energy released when the protons realign with the magnetic field. The diverse tissues in the body can be identified based on the energy released and the time required for realignment with the magnetic field [34] [35]. There are a variety of MRI techniques - structural, functional, diffusion, etc. - and these are described below

The **structural MRI** generates highly detailed three-dimensional anatomical images, although the specifics of various MRI sequences won’t be extensively covered in this thesis. A subsequent section will outline the particular sequences used in this study. As mentioned previously, in AD, neurodegeneration leads to the gradual loss of neurons and their connections in the brain, resulting in observable brain atrophy detectable via structural MRI.

To obtain quantitative results, software like freesurfer can be used. Freesurfer is a tool for the analysis and visualization of structural, functional, and diffusion neuroimaging [36]. For a structural MRI freesurfer is able to segment the brain using a specific atlas. An atlas is a map of the brain that labels specific regions. Different atlases exist, which segment the brain in various way [37]. With freesurfer a volumetric measure of different structures of the brain can be obtained. These measurements offer quantitative data regarding brain atrophy. 2 examples of atlases can be seen in figure 3, it represents the desikan-killiany atlas and the Destrieux

atlas.

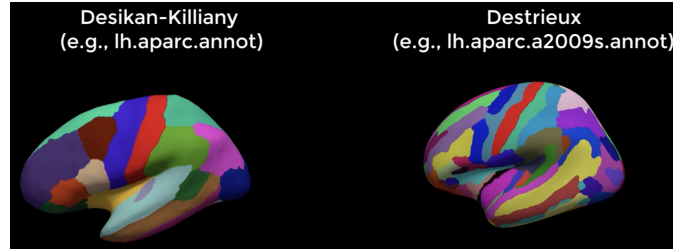


Figure 3: Example of atlas segmentation with the Desikan-Killiany atlas and the Destrieux atlas. This image is from a freesurfer tutorial [38].

The **functional MRI** is a method used to observe cerebral activity in the brain when the patient is in a resting state or performing a specific task [39]. The data can be analysed using a variety of software such as Freesurfer [36] or BrainVoyager [40]. One of the uses of functional MRI shows, for example, that in Alzheimer’s disease there are correlations between cerebral connectivity and the cognitive problems experienced by the patient. [41].

Lastly, the **diffusion MRI (DTI-MRI)** is a specific MRI technique that uses the diffusion of the water in vivo. The movement of the water is constrained by the cerebral micro-structures, and this movement is used to observe indirectly these micro-structures after processing with a model of diffusion [42]. This type of MRI allows researchers to reconstruct tracts (neural pathways). The goal is to observe the changes of these tracts through the evolution of the disease like in the present study [43]

1.5.3 Positron Emission Tomography Scan

The PET scan is a minimally invasive imaging procedure, it uses a small amount of radioactive material, typically ^{18}F -fluorodeoxyglucose (FDG), injected into the body. Various radioactive tracers are available, each binding to different molecules. Through this process, the visualization of these distinct molecule’s positions within the brain becomes possible. The radioactive tracer emits a positron (β^+) that will produce 2 photons (gamma rays) by annihilation when it encounters a nearby electron (its anti-particles). The emitted gamma rays travel in opposite directions and are captured by a sensor. Using this sensor, the emission’s location of the positron can be determined, enabling the creation of 3D images. This allows medical professionals to evaluate metabolic activity and organ functions [44]. Now,

an explanation on how this method is used to diagnose AD will be detailed.

Initially, in the absence of radioactive tracers binding to particular AD molecules, PET FDG was employed. This radioactive tracer is used to study the glucose metabolism in the brain. With this method, early brain metabolism changes preceding the onset of AD symptoms can be identified [45]. Currently, MRI and PET FDG imaging methods are used as valuable indicators of neurodegeneration. For example a volume loss in the temporal-parietal and/or hippocampal regions, might be a good clue in the diagnosis of AD. Nevertheless, these imaging techniques lack the ability to distinguish the specific characteristics of AD.

Next, medical professionals turn to PET amyloid imaging, which employs a different radioactive tracer to detect amyloid- β in the brain. Amyloid- β , as described in the pathological section, can be found in other diseases besides AD. While this biomarker offers improved precision compared to MRI and PET FDG, it remains insufficient for diagnosing AD based solely on a specific amyloid- β threshold. Another critical issue with amyloid- β is that many elderly individuals exhibit elevated amyloid- β levels yet maintain cognitive normality, indicating a lack of correlation between amyloid- β and cognitive impairment [46].

To achieve even greater specificity in diagnosing AD, the researchers turn to PET tau imaging. This PET scan employs a different radioactive tracer to identify the tau protein. The role of the protein tau in AD has been discussed in the pathological section. PET tau can detect tau pathology in the brain. Although not exclusive to AD, it offers a significantly higher level of specificity compared to amyloid- β . One notable advantage of tau protein detection is its strong correlation with cognitive impairment, when compared to amyloid- β detection [46]. In addition, the stage of Braak can be estimated using the PET tau [47]. This classification system categorizes patients based on neurofibrillary changes. These changes delineate distinct stages reflecting the advancement of AD [48].

Freesurfer is able to analyze the PET scan and gives the standardized uptake value ratios (SUVRs) for regions of the brain. The SUVRs is the most common way to obtain a quantitative value from PET scan. It is calculated by measuring the absorption of the radiotracer in an area of the brain compared to a reference region [49]. Freesurfer is also able to use the partial volume correction (PVC) to solve the limited resolution of the PET scan called 'partial volume effects' (PVE) [50][51].

1.5.4 Biomarker solubles

Doctors can also use a patient's cerebrospinal fluid (CSF) to detect amyloid- β and tau earlier than with a PET scan. However, this method comes with two drawbacks: the first one is that it's invasive since it requires a lumbar puncture. The second is that it cannot pinpoint the exact locations of amyloid- β or tau accumulation in the brain, unlike a PET scan, which provides the ability to localize the presence of these molecules in the brain. [52]

1.5.5 Challenges in Alzheimer's Disease Diagnosis

As described in this section, the diagnosis of AD is challenging due to the non-specific nature of many of the biomarkers that have been identified. These biomarkers often overlap with various other diseases. For example Lewy body disease, frontotemporal degenerations, or hippocampal sclerosis. This explains that there is currently no established threshold of biomarkers for diagnosing AD. Another important thing is the rate of progression of the disease. The cognitive impairment progression is really specific for each individual. For example a subject that has a subjective cognitive impairment are twice as likely to develop to the MCI stage and this evolution can take years to develop as well as never evolve to this stage. [53][54]

1.6 Treatments

Despite the lack of a cure for AD, there are several treatments available to manage symptoms and slow its progression. These treatments include both medications and lifestyle changes. While the disease remains debilitating, these options provide a way to improve patients' quality of life and potentially extend their functional abilities.

While some drugs are used only to alleviate the symptoms of AD, there are now two medications, *aducanumab* and *lecanemab*, which aim to slow down cognitive and functional decline. However, these medications are not curative and are only appropriate for individuals with early AD who exhibit evidence of amyloid- β buildup in the brain [55]. These two medications are still undergoing clinical studies.

In addition to drug treatments, non-drug therapies, such as physical activity, memory exercises, and music and art therapy, can help maintain or improve cognitive function and overall quality of life for Alzheimer's patients.

Overall, the current approach to AD is to slow down the progression of the disease and manage the symptoms experienced by the patient.

2 Algorithms

As described earlier, the goal of this thesis is to try different models to accurately predict when the patient will evolve to the dementia stage. Now that the presentation of the disease has been made, we will see the different algorithms that will be used to achieve this goal. The visualization algorithms will be presented first, followed by the prediction algorithms.

2.1 Visualization and dimension reduction Algorithms

2.1.1 t-Distributed Stochastic Neighbor Embedding

The t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm is a popular technique for dimensionality reduction and visualization of high-dimensional data. It aims to map the data points into a lower-dimensional space while preserving their local and global structure. All the explanations of this algorithm will be based on the article of Laurens van der Maaten and Geoffrey Hinton [56].

The first step is to compute the conditional probability. This probability represents the similarity between two points x_i and x_j from a high dimensional dataset.

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2 / (2\sigma^2))}{\sum_{k \neq l} \exp(-\|x_k - x_l\|^2 / (2\sigma^2))}$$

Here σ represents the variance of the gaussian.

The conditional probability can also be computed using y_i and y_j points. Here, y_i and y_j correspond to the reduced-dimensional representations of the original high-dimensional points x_i and x_j , respectively. This conditional probability will be denoted as q_{ij} . But instead of a Gaussian that has a short tail which will cause the crowding problem, the t-SNE uses a Student t-distribution with 1 degree of freedom. In summary it gives this formula for q_{ij} :

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

The main idea of the t-SNE algorithm is to minimize the mismatch between the two conditional matrices p and q . For this the algorithm minimizes a single Kullback-Leibler divergence between the joint probability distributions, P and Q , representing the high-dimensional and low-dimensional spaces respectively.

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

The goal of minimizing the Kullback-Leibler divergence is achieved by using a gradient descent method.

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j) (1 + \|y_i - y_j\|^2)^{-1}$$

2.1.2 Principal Component Analysis

The goal of Principal Component Analysis (PCA) is to reduce the dimensionality of a high-dimensional dataset while retaining as much information as possible. It achieves this by finding a new set of orthogonal axes called principal components that capture the maximum variance in the data. The first principal component explains the direction of the largest variance, the second principal component explains the direction of the second largest variance, and so on. By selecting a subset of the principal components, the data can be represented in a lower-dimensional space while minimizing the loss of information [57]. The figure below 4 is an example of the PCA execution on a 2D data. The projection space is also 2D but the axis of the graph changes to become axis with the highest variance in the dataset.

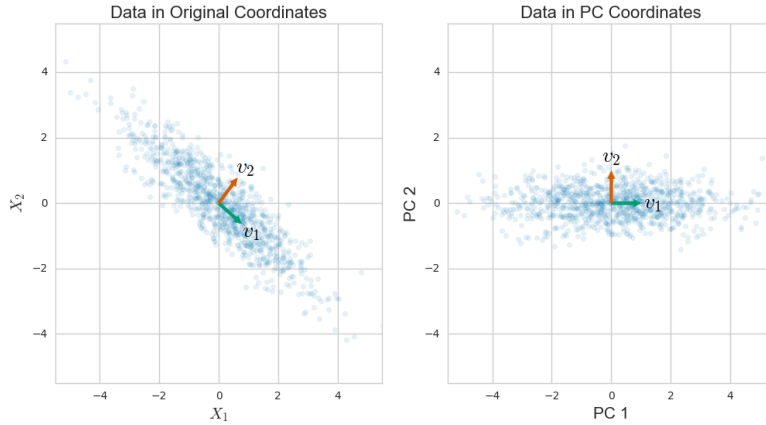


Figure 4: Example of the execution of PCA on data detailed in [58] image from [58].

Before discussing the algorithm of Principal Component Analysis (PCA), it is essential to emphasize that data standardization is a prerequisite.

To explain the PCA let's take a matrix X with the shape $m \times n$. The first step of the PCA algorithm is to compute the covariance matrix. Here is a quick reminder of the covariance formula:

$$\text{Cov}(X_1, X_2) = \frac{1}{m-1} \sum_{i=1}^m (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$

So the covariance matrix in the end will look like this :

$$\mathbf{C} = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Var}(X_n) \end{bmatrix}$$

To compute the eigenvalues of the covariance matrix $\mathbf{C}$, various methods can be employed, such as using the eigenvalue formula. One common approach is to solve the characteristic equation:

$$\det(\mathbf{C} - \lambda \mathbf{I}) = 0$$

To find the eigenvectors corresponding to the eigenvalues, the formula below can be used: [57]

$$\mathbf{C}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}$$

Here, $\mathbf{C}$ represents the covariance matrix, $\mathbf{V}$ is a square matrix with its columns as the eigenvectors, and $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues.

Now, it is necessary to perform data whitening. The primary objective of whitening is to eliminate any correlation or redundancy within the data, guaranteeing that the principal components derived through PCA are uncorrelated and have uniform variances. [59]

$$\mathbf{W}_{\text{PCA}} = \mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}^T$$

Now, let us explore ways to reduce the dimensionality of the dataset. The points from the high-dimensional space must be projected into a lower-dimensional space. To determine the number of dimensions, there are two options: specify the desired dimensionality or specify the percentage of variance to retain. This percentage is calculated as follows:

$$\text{Percentage of Variance} = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^n \lambda_i} \times 100$$

There is an alternative method for performing PCA known as singular value decomposition (SVD). However, the details of this method will not be covered in this thesis.

2.2 Predictive Machine Learning Algorithms

Before diving into the models for machine learning some preparation needs to be done on the data. A standard way to use the data correctly for machine learning is to first split the data in two to get a training set of the data and a test set to try to predict and get the score of the model. A standard split is to take between 20 to 30 % of the data for testing. Then the training set needs to be split again because models can be tuned some parameters called hyperparameters to improve the performance of the model. When fixing these parameters and directly evaluate them on the test set, there's a risk of optimizing them specifically for that test set. While this might yield excellent results for the initial test set, using these parameters on a different test set could result in poorer performance. This situation is commonly referred to as overfitting [60]. To prevent this a validation set is needed. In figure 5 a more graphical representation of these splits can be observed.

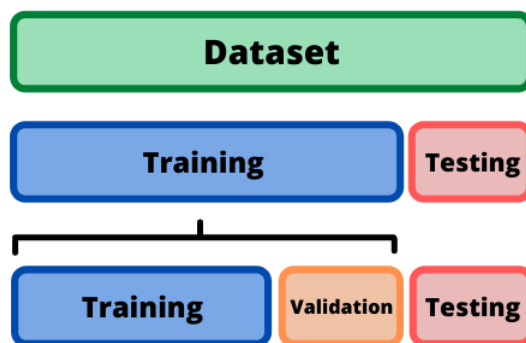


Figure 5: separation of the data for machine learning purpose.

This process can be further enhanced by employing k-fold cross-validation. This method involves splitting the data into k subsets, using each fold as a validation set while the rest act as the training set as you can see in Figure 6. This method minimizes the bias because it uses several validation sets and with that the model will not be as influenced by the specific cases. Furthermore, this method optimizes the use of the data by using all the training data for training and validation. In addition, this method provides a more robust performance evaluation than using a single validation set because it allows the model to be evaluated over multiple training iterations [61].

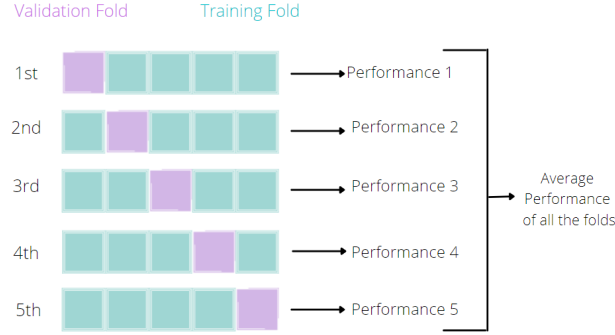


Figure 6: illustration of the split of the data for a kfold methods.

Even so, models tend to favor the majority class, as this improves their results. This problem is known as class imbalance. To solve this, an effective technique is to use stratification. This method ensures that each fold has the same class distribution (in the training, test and validation folds).

2.2.1 Support Vector Machine

The core idea of SVM is to find a hyperplane (a line in 2D, a plane in 3D, and a higher-dimensional equivalent in higher dimensions) that best separates two classes of data points. This hyperplane should maximize the margin, which is the distance between the hyperplane and the nearest data points of each class. The data points closest to the hyperplane are called support vectors. [62] All the terms that have been described here are illustrated in Figure 7. The figure illustrates a small example of classification between 2 classes + and -.

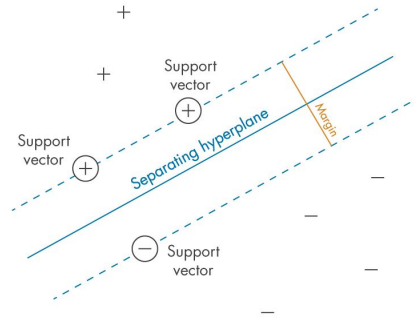


Figure 7: Here is an example of an SVM. It showcases various discernible features, including the separating hyperplane, which effectively divides the two classes (label + and -), and the margin, representing the distance from the hyperplane to the closest instances of each classes. Adapted from [63].

As previously described, the goal of the SVM is to find the hyperplane that will separate the classes. The hyperplane is defined by the function below:

$$\langle w, x \rangle + b = 0$$

Where w is a vector perpendicular to the hyperplane, x is a point of the hyperplane and b stands for the bias.

The SVM wants to find the maximal margin "M" where the margin can be calculated as :

$$M = \frac{2}{\|w\|}$$

Maximizing the margin can be done by minimizing the norm of the vector " w " while ensuring correct classification of the data points through constraints.

$$\begin{aligned} & \min \langle w, w \rangle \\ \text{Subject to: } & y_i(\langle x_i, w \rangle + b) - 1 \geq 0 \end{aligned}$$

In practical scenarios, it is often beneficial to relax this constraint slightly, allowing for some data points to be misclassified. This approach is known as the "soft margin classifier." Now if the problem is not linearly separable in the original feature space which is often the case in the real world. SVM employs a technique called the "kernel trick". This mechanism maps the data into a higher dimensional space using various kernel function. Commonly used kernel functions include the linear kernel, polynomial kernel, and radial basis function (RBF) kernel. The goal of the SVM is to find a hyperplane that separates linearly the classes in this higher dimension[64] [65].

2.2.2 Multilayer Perceptron

The Multilayer Perceptron (MLP) is a neural network used in machine learning and deep learning. The MLP is composed of artificial neurons layers. This algorithm can be used for a wide range of tasks, such as classification or regression problems.

The information in the MLP flows in a feed-forward manner. The information is going through each of the layers of the MLP. There are 3 different types of layers. The first layer is called the input layer. The number of neurons in this layer corresponds to the dimensionality of the feature space in the problem. The function of this layer is to pass the information to the next layer without any processing on the input. The second type of layer is the hidden layer. The role of this layer is to progressively process the data. The last layer is called the output layer. This layer is the prediction or decision of the model [66]. The architecture described here can be observed in Figure 8.

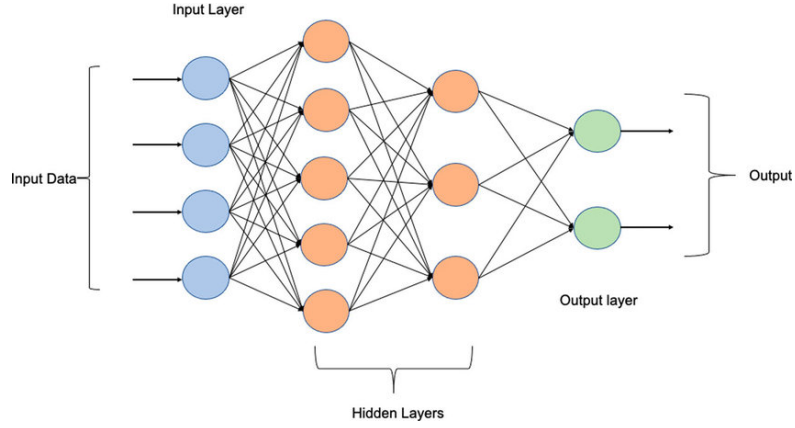


Figure 8: This figure illustrates the architecture of an MLP. The input layer can be seen in blue it has the same number of neurons as the dimension of the feature space. Then, in orange, are the hidden layers: the number of neurons and layers can be tuned to improve the performance of the model. Finally, in green, is the output layer which is the prediction of the model. Adapted from [67].

As detailed above, the layers are composed of artificial neurons. These neurons are also called perceptrons. Each perceptron receives input values. The perceptron makes a weighted sum over all of them and then adds a bias. Once it's done it passes the result into an activation function. This output is then passed to the neurons of the next layer and used as input value. A graphical illustration of this can be seen in Figure 9.

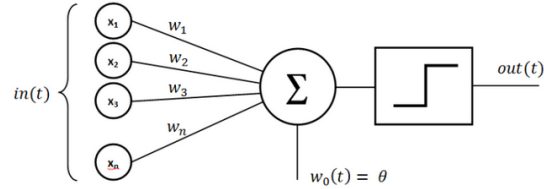


Figure 9: Here is a graphical representation of the functioning of a perceptron. The input values, denoted as x_1, x_2 , etc., undergo weighting by their respective parameters w_1, w_2 , etc. The perceptron aggregates these weighted inputs and processes them through an activation function, shown here as an S-shaped curve. Adapted from [68].

Activation functions are nonlinear and must remain so for non-linear problems. This is because the composition of linear functions remains linear, and non-linearity is required to capture complex relationships in the data. This will allow the MLP to learn more complex patterns and relationships in the data. Common activation functions include sigmoid, hyperbolic tangent ($\tanh$), and rectified linear unit (ReLU) [69].

For the first iteration of the training, the weight and bias are set to small random values. Then, the data is fed through the neural network in a forward propagation process. The process of the neuron, described earlier, is iteratively repeated layer by layer until the output layer, which is the prediction of the MLP. Once it's done, the network will compute the difference between the prediction and the true output (supervised learning). This error is then propagated backwards through all the layers and each weight and bias will be updated to reduce the error of the prediction. The process of updating the network weights is called backpropagation [66].

3 Existing research

This section will presents several articles that share similar ideas or objectives to the research pursued in this thesis.

The first article presents an automated method to diagnose AD. This study uses an SVM with the MRI scans of the whole brain to differentiate between AD patients and elderly control subjects. The study is composed of Sixteen AD patients (average age 74.1 years) and 22 elderly controls (average age 72.3 years). The MRI images were segmented into regions of interest (ROI) and the SVM used the gray matter of the ROI to classify the patient. This method achieved a 94.5% average correct classification rate for AD and control subjects, with 96.6% specificity and 91.5% sensitivity [70].

This first study aims to classify whether the patient has AD or not, but the goal of this thesis is to predict the onset of dementia. Therefore, some studies with the same goal were found to compare with the results of this study.

In a systematic review, among the 116 studies examined most of them have their data from the Alzheimer’s disease neuroimaging initiative (ADNI). And most of them use either a SVM or a convolutional neural network (specific structure of an MLP) which in average gives a precision of 75.4% for the SVM and 78.5% for the neural network. However, many studies are trying to differentiate controls from AD patients rather than the progression to dementia, so these percentages are not really precise to quantify the prediction of the onset of dementia [71].

One study attempted to predict the onset of dementia. For this purpose they wanted to understand the progression of AD by analyzing longitudinal biomarker data from a large group of participants. Using a Bayesian model, they aligned the data to create a timeline of how biomarkers such as CSF levels, brain volume, and cognitive measures change over time leading to AD. The findings highlighted early changes in certain biomarkers and memory. They then tried to predict the onset of dementia for patient with AD by using a linear regression model. To evaluate the model they computed the root mean square error between the predicted and observed onset ages. In the end they achieved a mean error of 1.5 years [72].

Materials and Methods

4 Materials

In this section we will discuss the data collected from the doctors and used in this thesis. We will focus on the format of the data and the strategies employed to store it in a manner that facilitates the application of machine learning methods. It is essential to highlight that the initial goal was to merge files from two different studies into one comprehensive database and integrate it into a website for subsequent machine learning predictions. However, due to the considerable scope of this endeavor, it was deemed impractical for a single master's thesis. Consequently, it was decided to focus solely on the machine learning project, using only relevant data. Nevertheless, some progress was made towards the original goal of creating the database, and this work will also be presented in this thesis.

4.1 Presentation of the studies

4.1.1 The biomarkers study

The biomarker study is a multimodal and longitudinal study focused on AD. The goal of this study is to examine the disease progression within a patient cohort. Spanning over an 8-year period, it incorporates data collected from diverse sources. The study cohort comprises 153 individuals, including 31 controls, 87 with mild cognitive impairment (MCI), and 35 with subjective cognitive impairment (SCI). Among these participants, there are 81 women and 72 men, with an average starting age of 71.1 years at the commencement of the study. The graphs below 10, depict key demographic characteristics and distribution within the study cohort.

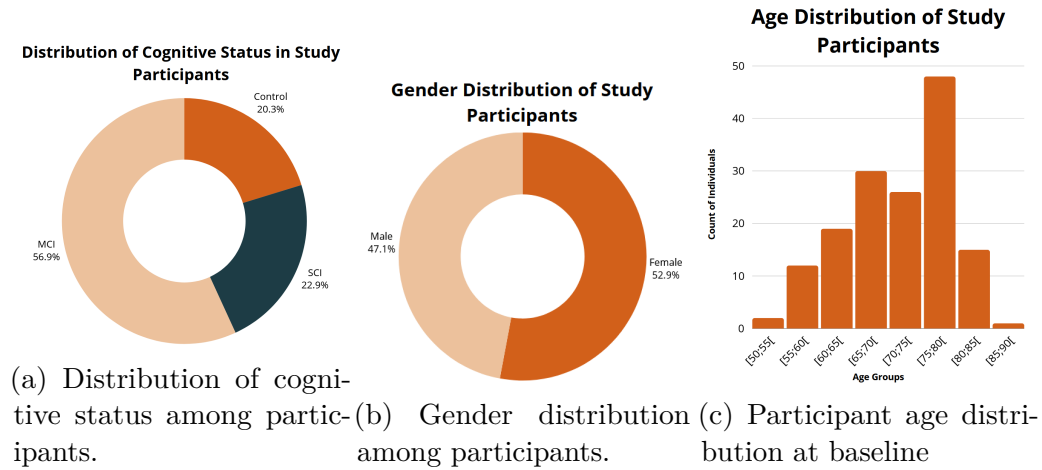


Figure 10: Demographic characteristics of the Biomarker study

This study is, as mentioned earlier, multimodal which means that it contains data from different sources. At the beginning of the study the data collected is composed of :

- Neuropsychological tests

MRI exam containing:

- Anatomical MRI (sequence used : 3DT1)
- Resting state

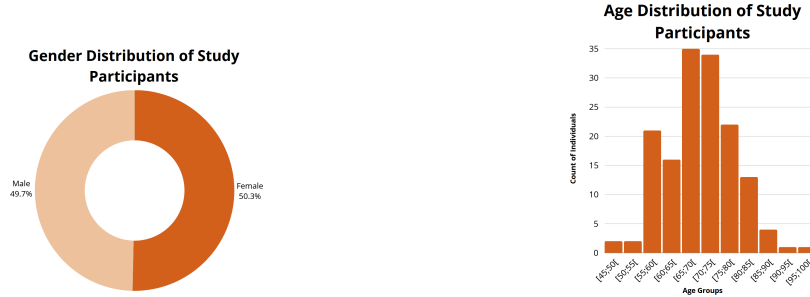
PET exam containing:

- PET FDG
- PET amyloid

After this initial stage, the patient has to come back every 2 years for the neuropsychological test. These tests are : MMSE, FCSRT, Luria, TMT, Naming lexis test, Fluency test, clock test, CERAD. These tests allow the psychologist to assess if the patient has reached the stage of dementia or not. For the anatomical MRI, one sub-group of the patients had to undergo an examination at the beginning of the study and 3 years later. While another segment of the cohort underwent MRI scans at the beginning of the study, and subsequently at intervals of 6 months, 1 year, and 1.5 years. For the PET scan there is only one at the start of the study for the two modalities.

4.1.2 The TAU study

The TAU study is also a multimodal and longitudinal study focused on AD. However some source of data is changed or added compared to the first study on biomarkers. The goal of this study is the same as the last one, to study the AD progression within a patient cohort, however, there are differences in the types of data collected between these two studies. Spanning over an 4-year period and currently ongoing, it includes data collected from several sources. The study cohort comprises 151 individuals, including 76 women and 75 men, with an average starting age of 69.4 years at the beginning of the study. The Figure 11, shows key demographic characteristics and distribution within the study cohort. As mentioned above, this study is still in the recruitment phase and these statistics are therefore taken from the Excel neuropsychological test of January 2023.



(a) Gender distribution among participants. (b) Participant age distribution at baseline

Figure 11: Demographic characteristics of the TAU study

This study is, as mentioned earlier, multimodal which means that it contains data from different modalities. At the beginning of the study the data collected is composed of :

- Neuropsychological tests

MRI exam containing:

- Anatomical MRI (sequence used : 3DT1, T2, flair)
- Resting state
- DTI-MRI

PET exam containing:

- PET amyloid (on the control patients or when there is a doubt on the result of the lumbar puncture)
- PET TAU

Blood and CSF:

- Lumbar puncture (for patients but the controls have the choice between a lumbar puncture and a PET amyloid)
- blood test

After this initial stage the patient has to come back 2 years later for theses series of tests :

- neuropsychological test

MRI exam containing:

- Anatomical MRI (sequence used : 3DT1, T2, flair)
- Resting state
- DTI-MRI

PET exam containing:

- PET TAU (Only for patients who have their initial PET TAU in a stage of braak between 1 to 4 because they are highly susceptible of evolving.)

After four years into the study, the patients return only for the neuropsychological exams. Additional tests may be added as the study progresses; currently, only the initial patients have reached this step.

The neuropsychological exams contains these tests : MMSE, TICS, FCSRT, Luria, TMT, Naming lexis test, Fluency test, clock test, CERAD, ECog, Logical memory test from the Wechsler Memory Scale, The shape test from Doors and People tests. These tests allows the psychologist to assess if the patients have reached the stage of dementia or not.

For the machine learning prediction, we have chosen to concentrate solely on the Biomarker study. The reason behind this decision is that the TAU study is relatively recent. This could lead to false negatives in the predictions, making the analysis less accurate. Moreover, the study has different modalities so we can't use the patients in the tau study who have already reached dementia because the data collected is not the same as in the biomarker study. As a result, we opted to focus on the Biomarker study, which already spans an extensive 8-year follow-up period, and provides more reliable data for the machine learning analysis.

4.2 MRI

As presented in the State of the Art section, Freesurfer is a powerful open-source software package designed for the analysis and visualization of neuroimaging data. Freesurfer is used with 2 of the 3 sequences made from the 2 studies described earlier. This software goes through a series of computational steps to analyse the 3DT1 sequence to compute all the measures of the brain (e.g. : area, volume,...) and the T2 sequence is used in combination with the 3DT1 to compute the metrics of the hippocampus and the Amygdala (only for the tau study since biomarker don't use a T2 sequence). At the end, Freesurfer generate several output text files.

The Figure 12 synthesizes the processing step to go from the MRI image to the final csv file. Freesurfer takes the anatomical MRI of a patient and generates multiple text files. An in-house software will then take these text files and turn them into multiple csv files. Each csv file will contain data from all patients corresponding to a particular brain region or segmentation of the brain based on an established brain Atlas.

For a detailed version about the work done to process the text file of freesurfer into the csv file we invite the reader to look in the appendix in the section 12.1 named "MRI Processing".

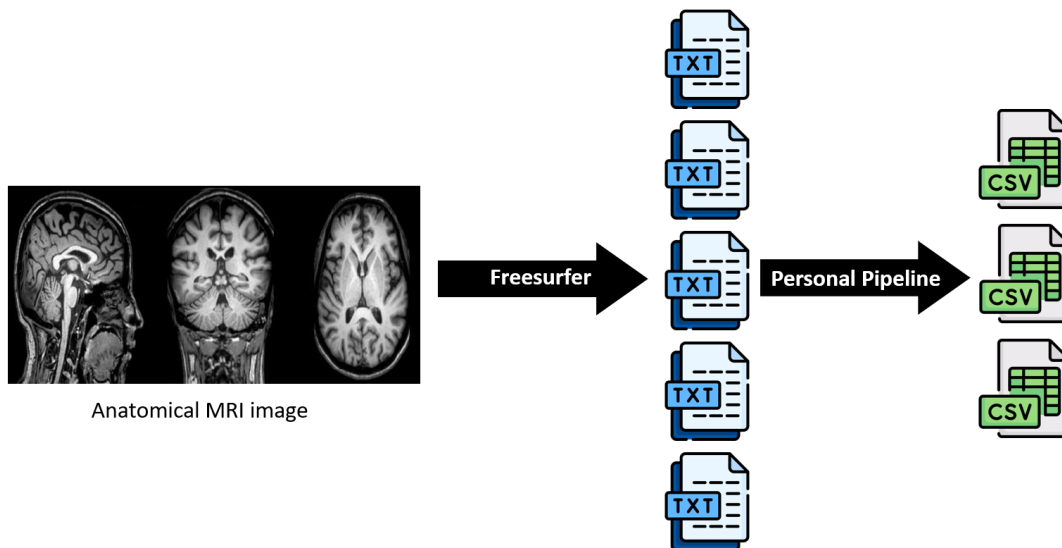


Figure 12: The figure provides a visual representation of the data cleaning process, outlining the sequential steps necessary to prepare the data for inclusion within a database. The anatomical MRI image is credited to [73].

As mentioned earlier, the 2 studies have collected a resting state (fMRI) as well as a DTI-MRI for the tau study but in the context of this thesis we didn't use them even in the database these data are not present. To see a global view of what is in the database, please refer to the section 5.5.

For our machine learning model, we made a deliberate decision to work with a subset of the available csv files. Specifically, we chose to focus on the general information about the brain, the hippocampus data, the subcortical data, and the Desikan Killiany data. These particular datasets were carefully selected as it gives an overview of all the brain and a specific zoom on interesting regions for AD (e.g.: hippocampus).

4.3 PET scan

As explained in the State of the Art, the PET scan is a medical imaging technique used to assess physiological functions. Freesurfer can be used to extract the data from the PET scan. Different molecules have been studied for AD. Therefore different PET scan are able to localize and quantify them. In this thesis, we focused on 2 PET scans the PET FDG and the PET Amyloid. The PET tau hasn't been used due to the study's current lack of sufficient longitudinal duration of the study, as previously mentioned.

The Figure 13 summarizes the steps to go from the images of the PET scan to the final csv. Freesurfer uses the images of the PET scan to produce 2 files for each of the PET scans. The first one gives the SUVRs without the PVC correction and the second one contains the SUVRs with the PVC correction. The reference region for the SUVRs is the pons. The format of the file is changed for a CSV format for convenient future use.

As for the MRI the steps to go from the text files to the csv files will be detailed in the appendix section 12.2 named "PET scan Processing".

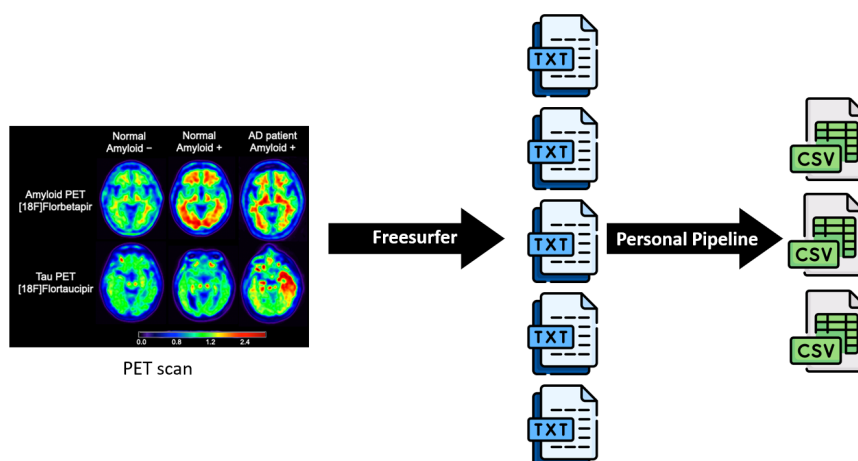


Figure 13: The figure provides a visual representation of the data cleaning process, outlining the sequential steps necessary to prepare the PET scan data for inclusion within a database. The PET scan image is credited to [74], the illustration showcases PET amyloid and TAU scans for three distinct patients: one classified as normal, displaying low amyloid and tau accumulation; another healthy patient exhibiting high accumulation of amyloid and tau; and the last patient, diagnosed with AD, with elevated levels of both amyloid and tau.

4.4 Cognition test

Throughout both studies, patients who volunteered to participate were scheduled for cognitive assessments with a neuropsychologist every two years. This periodic evaluation allowed researchers to monitor any potential changes or progression of neurodegenerative conditions. The neuropsychologist evaluates dementia by analyzing the results of these cognitive tests, providing valuable insight into the patients' cognitive functions. All the tests done in these studies have been described in the state of the art section.

The neuropsychologist stores the test results in an Excel spreadsheet, containing the values of the tests for all the patients and their respective time steps. After discussion with the neuropsychologist, we decided to split this file into 4 files based on the test categories and save them in 4 output csv. This division ensures a more organized and focused approach to handling the data, making it easier to perform specific analyses on each test category. For the TAU study, an additional file is included. As part of this thesis, a program was created to divide the original excel file into several files based on the categories of each test. A visual representation of this process have been illustrated in Figure 14.

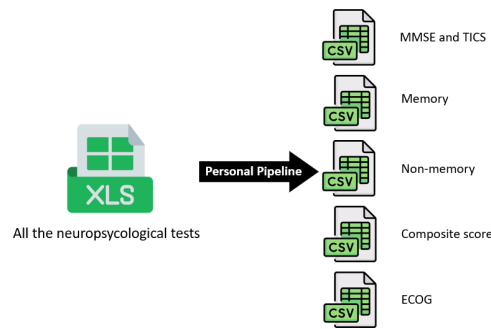


Figure 14: This illustration represents the pipeline to split the excel file of the neuropsychological tests into sub-files with different categories

One of the csv files contains the MMSE and TICS of the patients for all the time steps that the subject has completed. These two tests are good to get a general idea about the state of the patient's cognitive function over time.

Another file called "Memory" contains the results of the tests that are memory-related. In these studies, we stored the result of the FCSRT test, the logical memory test from the Wechsler Memory Scale and the results of the shape test form Doors and People tests.

Additionally, the 'Non-Memory' file compiles outcomes of non-memory-related tests, including the Luria test, TMT test, lexis naming test, fluency test, clock test, and CERAD test.

Another csv file contains the patient composite score, calculated from all the tests and categorized into memory, executive function, visuospatial, and language. This file also contains the date of each test and the patient's z-score for each test. The z-score expresses the deviation from the mean in standard deviations.

The last csv file contains the result of the ECOG test, which is only performed in the TAU study.

4.5 Database

This section will introduce the database created for this thesis. The database architecture is structured into five folders: Demographic, Neuropsychological tests, MRI, biomarker, and Plasma. Each folder will be explored in detail, providing an explanation of its contents and organization. The Figure 15 illustrates the architecture of the database.

```
> Demographic
> Neuropsychological tests
  > MMSE and TICS
  > Memory
  > Non-Memory
  > Composite score
  > ECOG
> MRI
  > general
  > desikan_killiany
  > sous_cortical
  > ...
> Biomarker
  > FDG
    > SUVR with PVC
    > SUVR without PVC
  > Amyloid
    > SUVR with PVC
    > SUVR without PVC
  > Tau
    > SUVR with PVC
    > SUVR without PVC
    > braak stage
> Plasma
```

Figure 15: Database Architecture

The first folder is the demographic folder. It contains general information about the patient (year of birth, id in the study he participated in, and a global id since this db regroups 2 studies).

The second folder regroups the results of the neuropsychological tests of the two studies. The data are stored into different csv that regroup the tests of the same categories as explained in more detail in the section 5.4.

Then, the folder for the MRI stores only the data from the anatomical MRI. It contains the csv created from the Freesurfer text file as described in the section 5.2.

The biomarker folder contains the 3 biomarkers that have been used in both of the studies. So there are 3 sub-folder on for each of the biomarkers (FDG, Amyloid, and TAU). The FDG and amyloid contains the data extracted from the Freesurfer file as explained in the section 5.3 called PET scan. The Tau subfolder contains the data from the Freesurfer text files and the stage of braak of the patient.

Finally, the plasma file contains the results of the blood analysis. These results include the amount of amyloid and tau in a blood sample.

At the beginning of this thesis 2 additional folders were planned. The first was for the CSF data, which should have included information on amyloid and tau in the CSF. The second, was a medical folder that was intended to gather broader health-related data about the patients, such as height, weight, blood pressure, and other general health metrics. However, this data has not been processed and the priority of the thesis shifted to machine learning instead of the database.

5 Methods

5.1 Inclusion criteria

Prior to constructing the models, the data was imported into the program using the pandas library [75]. Subsequently, unnecessary data were removed. Specifically, patients were excluded from the study if they had an insufficient number of time steps with the neuropsychologist, making it uncertain whether they had progressed to dementia or not. Patients with less than six years of follow-up were excluded, except for those who were already diagnosed with dementia within six years, in which case the follow-up for this study ceased after the onset of dementia. The biomarker study has some patients with 8 years of follow up. However, due to the limitations of the available data, we will focus on a 6-year period. Out of the 139 patients included at the start of the study, only 3 patients developed dementia after an eight years of follow-up. Patients who developed dementia after 8 years are classified as never having developed dementia.

Additionally more patients were excluded after the data visualization that will be described later in this thesis. This visualization step has allowed us to identify patients in the dataset who have reached the stage of dementia, but due to other diseases than AD. Since we want to focus on AD those patients have been excluded for the prediction models. A total of 34 patients have been excluded. 26 due to their lack of available timesteps and 8 due to an other disease than AD.

5.2 Preprocessing of the Data

Missing data is a common problem in machine learning, particularly when working with large dataset [76]. These missing values in the data need to be addressed. Upon closer inspection, it was found that there were numerous missing values in the neuropsychologist test data. Consequently, the decision was made to temporarily exclude these data from the analysis. However, in a subsequent phase, we dedicated time to address these missing values. Therefore, some tests have been carried out using the data from the cognitive tests. All missing values in the entire dataset were replaced by the median value of their respective columns, according to their labels [76].

Two methods of data normalization were used in this thesis to investigate their impact on the results of the models.

One commonly used method for normalization is z-score normalization [76]. It operates by subtracting the mean (μ) of each feature (column) in the dataset from its individual values (X), then dividing this difference by the standard deviation (σ) of that feature. This process ensures that the data is transformed to have a mean of zero and a standard deviation of one. This normalization can be expressed by the formula:

$$Z = \frac{X - \mu}{\sigma}$$

The second method of normalization used is the min-max normalization [76]. This technique rescales the data, constraining it within the range of 0 to 1. To achieve this, each data point is adjusted by subtracting the minimum value (min) of the respective feature and then dividing this result by the range, which is calculated as the difference between the maximum (max) and minimum values of that feature. Mathematically, the min-max normalization can be represented by the formula:

$$\text{Normalized value} = \frac{X - \min}{\max - \min}$$

5.3 Visualisation of the data

Before we dive into machine learning models, we want to plot the data to see if a pattern emerges. We need to reduce the dimensionality of the data to 2 or 3 so that it can be plotted on a 2D or 3D graph. This reduction is accomplished using the t-SNE or PCA algorithms described in the State of the Art chapter. After applying the procedure described in Section 5.2, the t-SNE and PCA functions are implemented using the sklearn library [68]. First, the processing is performed without the Cognition data. Two 2D and one 3D graphs are generated by t.SNE, while PCA generates two 2D and two 3D graphs. Finally, two more 2D graphs are generated with all the data by t-SNE and PCA, respectively.

5.4 Machine learning models

In the "State of the Art" section, we introduced two models: SVM and MLP. This section describes their implementation within this thesis, detailing how various methods, including testing different MLP architecture or adjusting hyperparameters for both MLP and SVM, were evaluated, considering whether dimension reduction was applied or not.

5.4.1 Dimension reduction

To see the impact of a dimension reduction algorithm on our model it has been decided to use the PCA. In order to understand its impact, a comparison was made between runs with and without PCA. When PCA isn't applied, our model receives the preprocessed data directly. However, in instances where dimension reduction is desired, the PCA algorithm from the scikit-learn package [77] has been used. To determine the number of dimensions retained by PCA, a fixed threshold based on the dataset's variance is applied. Typically, a variance range of 70% to 95% is recommended for this purpose. The variance can be seen as a measure of the amount of information retained from the original dataset. This can be summarized as a higher variance means more information is retained. However, seeking higher variance results in working with increased dimensions.

In figure 16, the red curves show the evolution of the variance as a function of the number of components retained. The blue bars show the variance of the component. They are sorted by decreasing order of variance. In green is the threshold that we have fixed for this thesis, we always keep 90% of variances. There is a graph for all the different modalities of data.

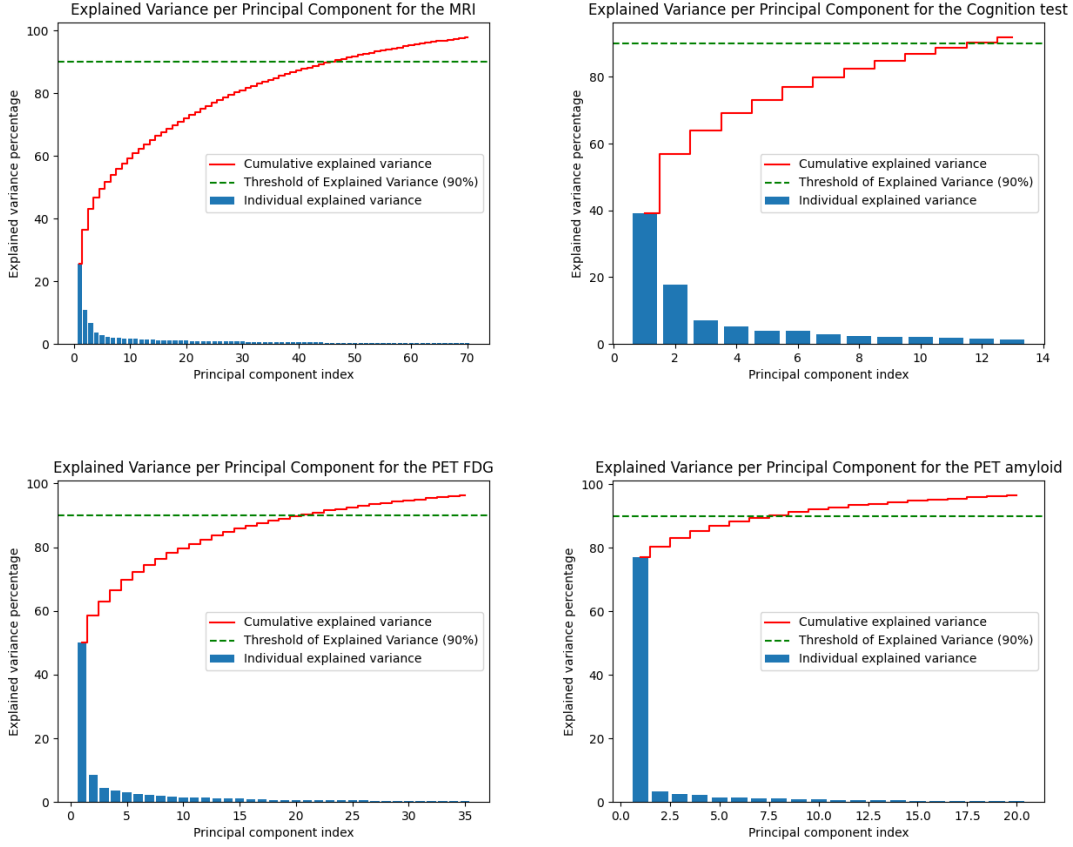


Figure 16: The red curve illustrates the relationship between the variance and the number of retained components. Meanwhile, the blue bars represent the variance of each component, sorted in descending order based on their variance. The graph on the upper left shows this for the MRI dataset while the one on the upper right shows it for the Cognition dataset. The graph on the lower left shows this for the PET FDG dataset while the one on the lower right shows it for the PET amyloid dataset. The green line at 90% is to represent the threshold of explained variance that we keep in this thesis for the different modalities of data.

5.4.2 SVM

Various tests were conducted using Support Vector Machines (SVM). Initially, an implementation of a binary classifier aimed to distinguish patients who would progress to AD and those who would not. This analysis was accompanied by a Receiver Operating Characteristic (ROC) curve to evaluate its performance.

Subsequently, a comparison between two normalization methods was undertaken to determine their respective efficacy. This comparison involved the creation of two confusion matrices, each derived from the results of 10 tests.

Additionally, another test was performed to assess the model's performance using different data combinations. Specifically, the evaluation aimed to ascertain if the model performed better with data solely from MRI and both PET scans or if its accuracy improved when including data from MRI, both PET scans, and cognition tests. To further evaluate the results, a confusion matrix was generated from 10 runs.

Then, a precision-recall curve was plotted, specifically utilizing the z-score normalization technique and use the data from the MRI, both PET scans and the cognition tests.

Finally, 2 tests were made to see the effect of dimension reduction. The first test is done with all data except for cognition, while the second test is done with all data. To evaluate the results, a confusion matrix was generated from 10 runs.

The SVM have been implemented using the scikit-learn library [77]. To optimize the hyperparameter of the SVM a grid search and a stratified kfold have been made using the scikit-learn package [77] for all the tests.

5.4.3 MLP

Two different architectures of the network have been made to see the impact on the prediction. The first one is a big network that takes the data as a whole. It means that all the data are in the same network from the start. The alternative architecture involves employing 4 distinct networks: one for the MRI data, one for each of the PET scans and one for the cognition when the cognition data are used. Following this initial processing, the outputs from these individual networks are combined within another network to generate the final prediction For better understanding, a graphical representation of the different networks have been made below.

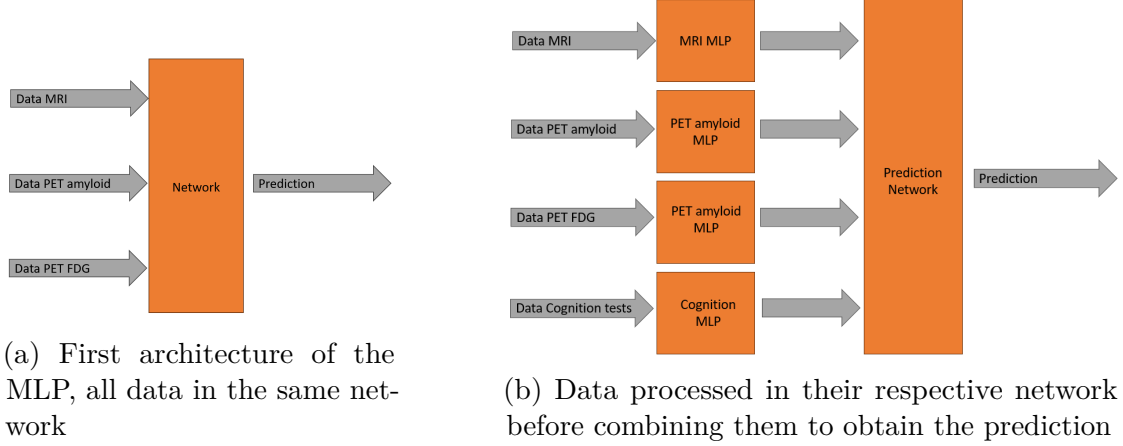


Figure 17: Architecture of the MLP used in this thesis

For the first architecture, a kfold was created to test different numbers of layers and different numbers of neurons in the layers. However, for the second architecture, due to difficulties encountered in implementing the kfold, the kfold was not implemented, instead only one validation set was used and the different hyperparameters were tested manually.

Four tests were conducted for each architecture. The initial two tests involved PCA for dimension reduction. The first of these tests used all data except for cognition data, while the second employed all available data. The subsequent two tests mirrored the aforementioned ones but were conducted without dimension reduction this time.

For the implementation of the MLP the tensorflow library was used [78]. For the first architecture, a stratified kfold was created using the scikit-learn package [77].

Results

6 Visualization

6.1 t-SNE

Below, you will find two scatter plots. The Figure 18 represents the t-SNE visualization of the data, where the labels of the data points have been altered. Patients who won't develop dementia are labeled as -1, while those who reach the dementia stage have their label changed from representing years before dementia to 1, indicating they will develop dementia. This modification helped better examine whether patient will develop dementia or not. On the other hand, the Figure 19 is identical to the first one, except that the labels remain unchanged. The labels indicate the time before dementia onset, and set at -1 for patients who never reach this stage. Both scatter plots are displayed to facilitate a straightforward visual comparison since the `random_state` parameter in the t-SNE function has been fixed, ensuring that the points are located in the same positions. The plot presented here displays the results obtained using the best hyperparameters identified.

As we can see in the figure 18, two clusters can be identified. However, it is important to note that some of the patient are in the wrong cluster. Even throughout all the hyperparameters that have been tested. We consistently observed that certain patient were always positioned in the middle of the wrong group.

Let's shift our attention to Figure 19, with this particular dimension and with the chosen algorithm, we observe that no distinct clusters can be identified. The labels representing the different year before the onset dementia in the patient population are not separated from each other as hoped.

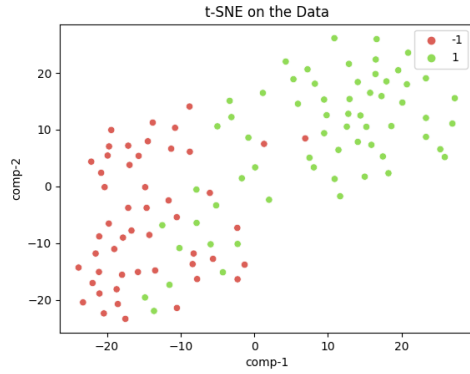


Figure 18: t-SNE in 2 dimensions on the data selected (MRI and both PET) and preprocessed, -1 means that the patient will not evolve to dementia, and 1 means they will evolve to dementia.

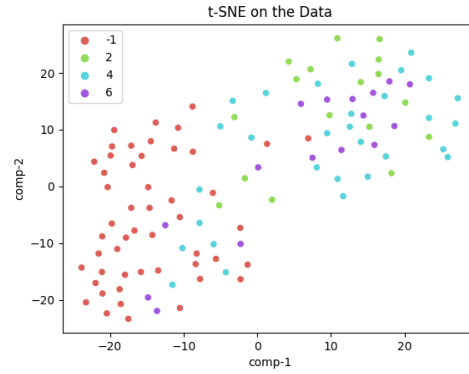


Figure 19: t-SNE in 2 dimensions on the data selected (MRI and both PET) and preprocessed, -1 means that the patient will not evolve to dementia, 2 means that they have evolved to dementia after 2 year, 4 is for for 4 years and 6 for 6 years.

The plot was also attempted in 3D to explore the possibility of obtaining additional information and achieving a clearer differentiation between the labels. However, the resulting plot turned out to be significantly more cluttered and challenging to comprehend. Even after trying to find the best parameter for the t-SNE function. This 3D plot can be observed in the figure 20. This is the best result achieved and the data is still in an agglomerate sphere with some outliers at the bottom of the figure.

t-SNE on the Data

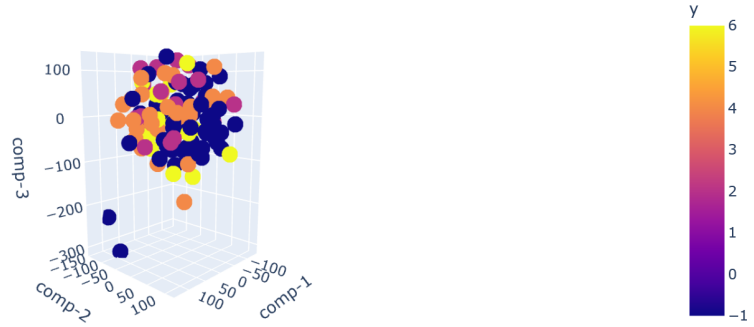


Figure 20: t-SNE in 3 dimensions on the data selected (MRI and both PET) and preprocessed, -1 means that the patient will not evolve to dementia, 2 means that they have evolved to dementia after 2 year, 4 is for for 4 years and 6 for 6 years.

After the missing values of the cognition dataset were filled. The data of the cognition tests were added in the plot with the data of MRI and both PET. Surprisingly, there is no more visible cluster on the figure 21, even with the modify labels described earlier.



Figure 21: t-SNE in 2 dimensions on all the data selected (MRI, both PET and cognition tests) and preprocessed, -1 means that the patient will not evolve to dementia, 2 means that they have evolved to dementia after 2 year, 4 is for for 4 years and 6 for 6 years.

6.2 PCA

Firstly we begin the visualization with 2 scatter plots in 2D of the data. For this we set the number of components in the PCA to 2 to be able to visualize the data in 2D. The 2 graphs are the same as in the t-SNE section. The Figure 22 has modified labels to see whether a patient will develop dementia or not. The Figure 23, on the other hand, don't modify the labels to see if clusters are formed by patient who reach dementia in the same year.

Once again, there are two clusters that can be distinguished in the figure 22. Once again, we can observe some patient in the wrong cluster. However, there are slightly fewer misplaced patient in this scatter plot.

If we look in the Figure 23, no clear distinct cluster can be observed for the patient that reach dementia in the same year, as we would have hoped.

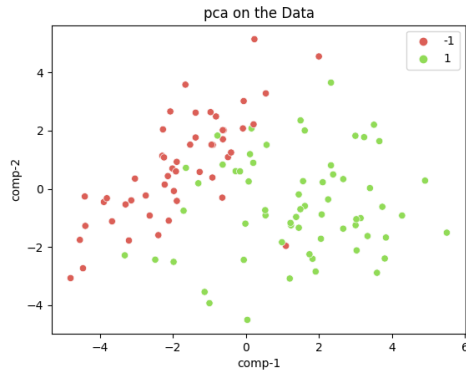


Figure 22: PCA in 2 dimensions on the data selected (MRI and both PET) and preprocessed, -1 means that the patient will not evolve to dementia, and 1 means they will evolve to dementia

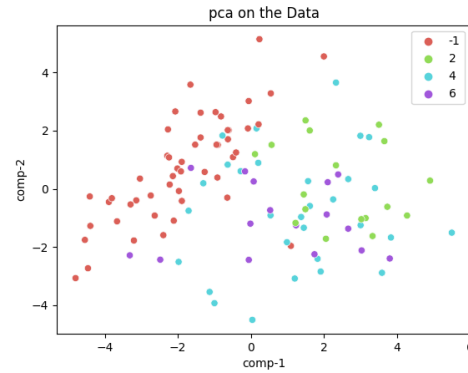


Figure 23: PCA in 2 dimensions on the data selected (MRI and both PET) and preprocessed, -1 means that the patient will not evolve to dementia, 2 means that they have evolved to dementia after 2 year, 4 is for for 4 years and 6 for 6 years.

The processed described earlier for the 2D plot is the same here but this time instead of taking 2 components from the PCA we take 3 components. This resulted in the creation of the following two graphs. The figure 24, illustrating the binary classification of patients, to see whether a patient will develop dementia or not. Then the figure 25, that don't modify the label to see if clusters are formed within the patients that reach dementia in the same year.

This time the 3D graph is much more readable than with the t-SNE algorithm. 2 clusters can really be seen in figure 24.

However, in figure 25, it's apparent that no distinct clusters can be identified. The labels representing the different years before the onset of dementia in the patient population aren't separated from each other, contrary to what was expected.

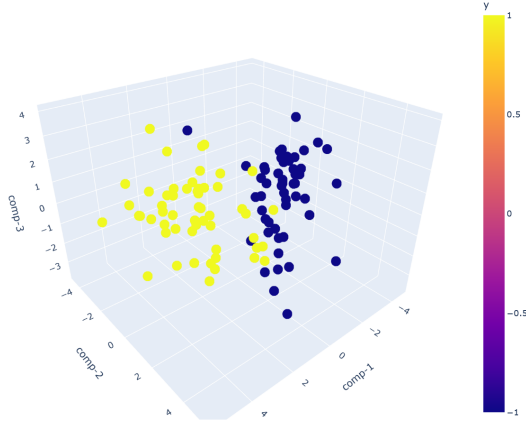


Figure 24: PCA in 3 dimensions on the data selected (MRI and both PET) and preprocessed, -1 means that the patient will not evolve to dementia, and 1 means they will evolve to dementia.

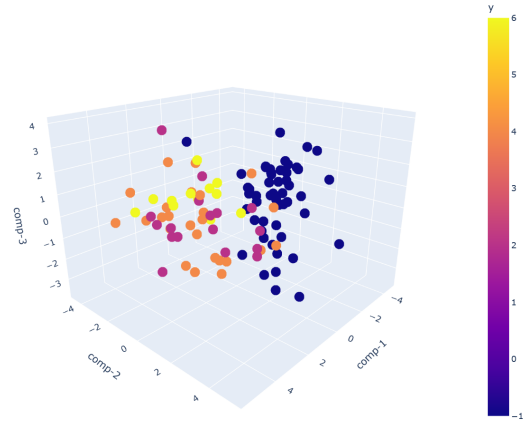


Figure 25: PCA in 3 dimensions on the data selected (MRI and both PET) and preprocessed, -1 means that the patient will not evolve to dementia, 2 means that they have evolved to dementia after 2 year, 4 is for for 4 years and 6 for 6 years.

Once the missing values in the cognition dataset were addressed, the cognition test data was incorporated into the plot alongside the MRI and both PET data. Surprisingly, despite modifying the labels to specifically highlight whether patients would progress towards dementia or not, there were no discernible clusters visible in Figure 26.

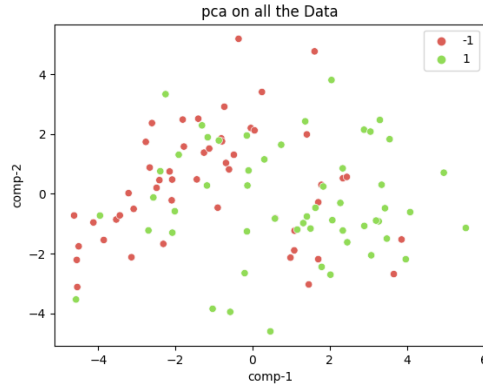


Figure 26: PCA in 2 dimensions on all the data selected (MRI, both PET and cognition tests) and preprocessed, -1 means that the patient will not evolve to dementia, and 1 means they will evolve to dementia.

7 Predictive Machine Learning Algorithms

7.1 SVM

As previously mentioned, the first test conducted was a simple classifier to determine whether the patient will develop dementia in the next six years due to AD. Two methods were used to achieve this goal.

The first method involved using SVM with data from both the MRI and PET scans. While, the second test included cognition data in addition to the components of the first test.

Out of 10 runs, the first test had a precision of 0.919 and a standard deviation of 0.0479, while the second test had a precision of 0.9524 and a standard deviation of 0.0301.

The ROC curve of one of the run for each test can be observed in figure 27 as well as their respective area Under the Curve (AUC).

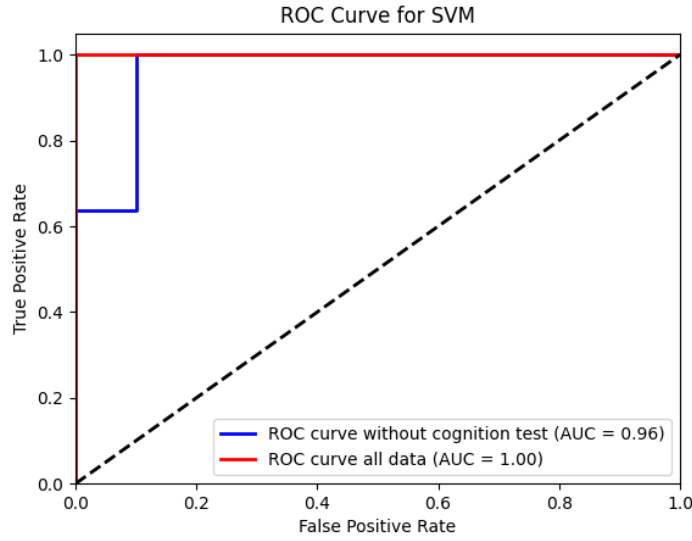


Figure 27: Two ROC curves represent a binary classification using an SVM. One curve depicts the classification with all available data, while the other excludes the cognition test. This binary classification aims to predict whether patients will develop dementia specifically due to AD.

Then a test was made to see the best method of normalization, but this time the classifier had to predict the onset of dementia. The data used for these tests included MRI, cognition tests, and PET scans. The SVM models with z-score normalization achieved a mean accuracy of 0.6381 and a standard deviation of 0.053 over 10 runs. However, the min-max normalization achieved a mean accuracy of 0.5714 and a standard deviation of 0.0673 over 10 runs. The cumulative confusion matrices for both methods are shown in the two graphs below. Figure 28 displays the results for the z-score normalization, while figure 29 shows the results for the min-max normalization. As we can see from the two combined confusion matrices, the two models have a tendency to overfit. This can be seen as the models predict most of the time that the patient will not progress to dementia or will progress to dementia after 4 years.

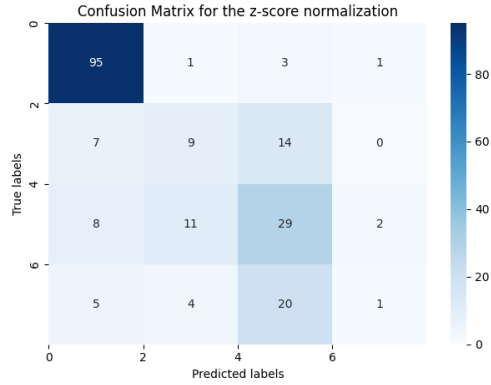


Figure 28: Combined confusion matrix of 10 runs to predict the onset of dementia. Prediction made by an SVM with the z-score normalization and all the data.

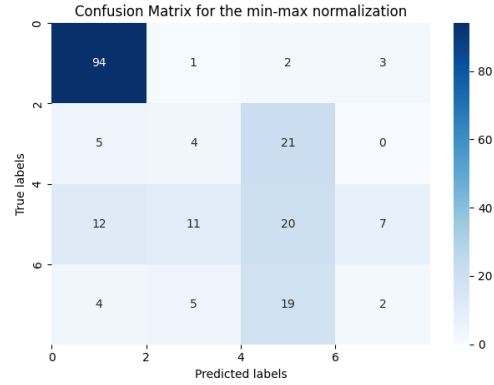


Figure 29: Combined confusion matrix of 10 runs to predict the onset of dementia. Prediction made by an SVM with the min max normalization and all the data.

For the next test, we tried predicting the onset of dementia with an SVM and z-score normalization, but this time without the cognition tests to see if they had an effect on the result. The achieved accuracy was 0.6238 with a standard deviation of 0.0497. Which is slightly lower than the score obtained with all the data. The combined confusion matrix of the 10 runs with the setup described above can be seen in figure 30. As for the two confusion matrices in figures 28 and 29, the models have a tendency to overfit. The model predicts either that the patient will not progress to dementia or that they will progress to dementia after 4 years. The model almost never predicts that the patient will progress to dementia after 6 years.

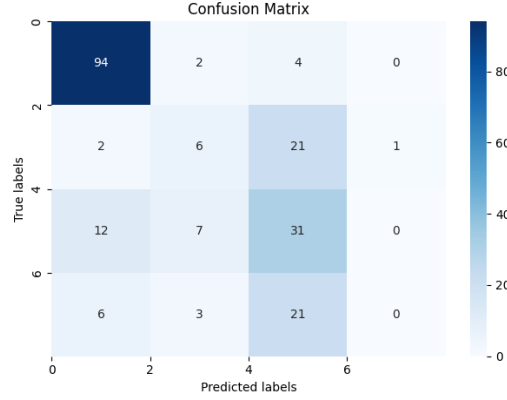


Figure 30: Combined confusion matrix of 10 runs to predict the onset of dementia. Prediction made by an SVM with the z-score normalization and all the data except the cognition test

A precision recall graph have been made in an iteration with all the data and a z-score normalization.

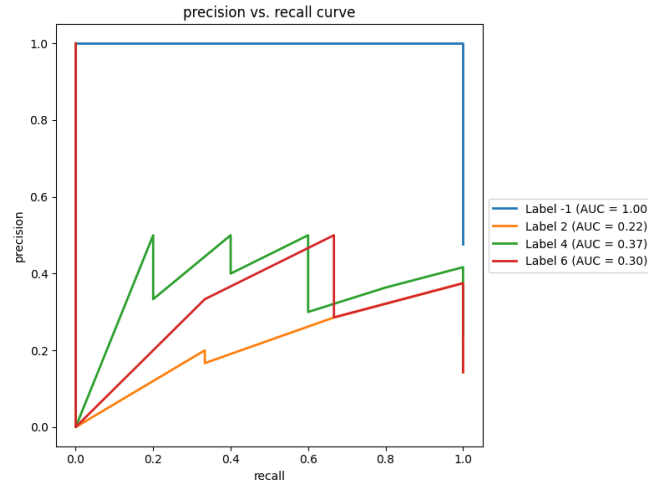


Figure 31: Precision recall graph of one run to predict the onset of dementia with an SVM the z-score normalization and all the data

The above tests were conducted using a PCA algorithm for dimension reduction. To assess the impact of this method on prediction accuracy, two additional tests were performed without PCA to compare with the previous result: one using all the data and another using all the data except the cognition data.

The initial test, conducted without PCA and using all available data except the cognition data, achieved a mean accuracy of 0.6095 with a standard deviation

of 0.0513 over 10 runs. In contrast, the second test, which excluded PCA but with all the data, achieved a mean accuracy of 0.6143 with a standard deviation of 0.0497. The combined confusion matrix of 10 runs can be found in figure 33 for the first test and in figure 32 for the second test. Even considering these parameters, the figures indicate a tendency towards overfitting in the model. It consistently predicts that patients will either not progress toward dementia or will do so after 4 years. The model rarely predicts that patients will develop dementia beyond the 6-year mark.

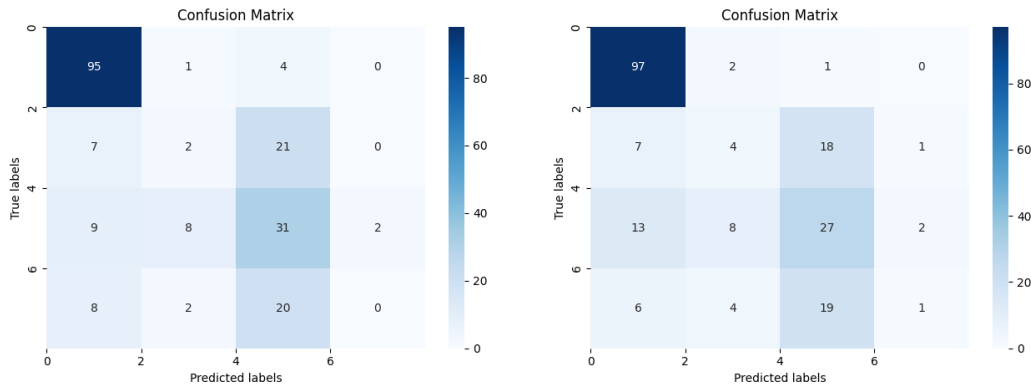


Figure 32: Combined confusion matrix of 10 runs to predict the onset of dementia. Prediction made by an SVM with the z-score normalization and all the data without cognition and without PCA.

Figure 33: Combined confusion matrix of 10 runs to predict the onset of dementia. Prediction made by an SVM with the z-score normalization and all the data without PCA.

7.2 MLP

As explained in section 5.4.2, two different architectures for the MLP have been tested. This section will present the results of the first architecture (only one network), followed by the results of the second architecture (5 networks).

7.2.1 One network

The initial test conducted using the single network architecture is a test to see the effect of the cognition data on the result. The first test uses all available data except for cognition, while the second test uses all available data. The first test achieved a mean accuracy of 0.6286 over 5 runs, with a standard deviation of 0.0467. The second test was slightly less precise, with a mean accuracy of 0.6095 over 5 runs and a standard deviation of 0.0762. The combined confusion matrix can

be seen in Figure 34 for the first test and in Figure 35 for the second test. Again, we can see in the 2 figures that even with a different algorithm the models have a tendency to overfit. It never predicts that the patient will progress to dementia after 6 years. And most of the time the model predicts that the patient will not progress to dementia or will progress to dementia after 4 years.

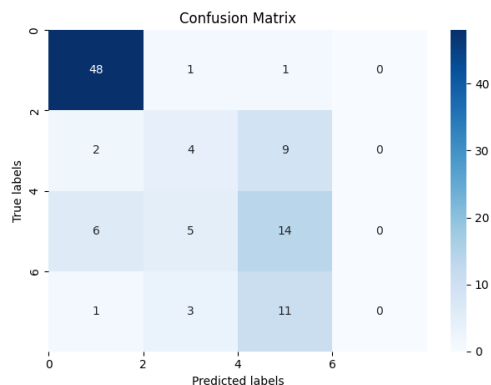


Figure 34: Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data without cognition and with PCA.

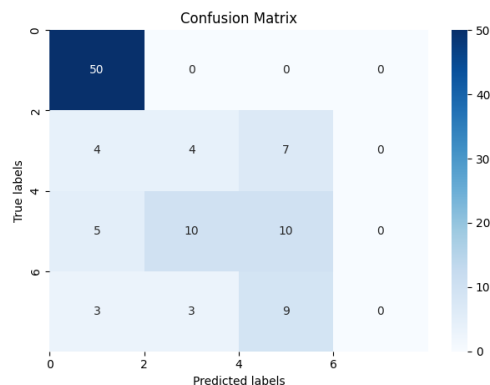


Figure 35: Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data with a PCA.

In order to see the effect of dimension reduction with PCA, two further tests were carried out. The previous 2 tests were done with dimension reduction. So this time it's the same two tests, but without the dimension reduction. As a reminder, the first test is done with all data except the cognition data and the second test is done with all available data. The first test achieved a mean accuracy of 0.619 out of 5 runs and a standard deviation of 0.0673. The second test has a mean accuracy of 0.6667 and a standard deviation of 0.0522. The combined confusion matrix of the 5 runs can be seen in figure 36 for the first test and in figure 37 for the second test. Once more, the figures highlight a persistent tendency towards overfitting in the models. They consistently fail to predict dementia progression after the 6-year mark. Moreover, most predictions lean towards either no progression towards dementia or progression occurring after 4 years.

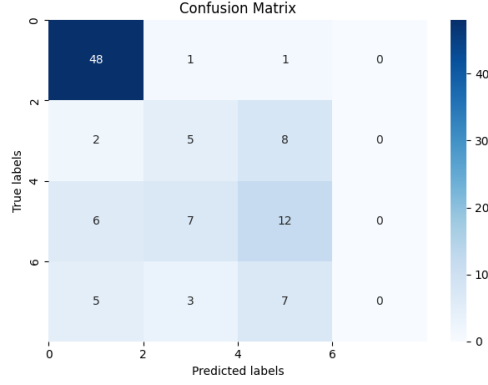


Figure 36: Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data without cognition and without PCA.

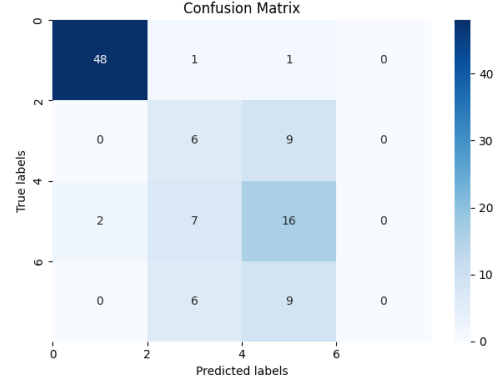


Figure 37: Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data without a PCA

7.2.2 5 networks

As mentioned previously, the second type of architecture consist of 5 different network showed in figure 17. As for the precedent architecture 4 tests were made.

The initial pair of tests aimed to assess how cognition data impacted the outcomes. In the first test, all available data except cognition was used, while the second test used all available data. Across 5 runs, the first test yielded a mean accuracy of 0.5524 and a standard deviation of 0.0646. The second test displayed a mean accuracy of 0.5524 over 5 runs and a standard deviation of 0.1111. Figures 38 and 39 illustrate the combined confusion matrices for the first and second tests, respectively. We can see from the 2 figures that with this neural network architecture the models have less tendency to overfit compared to the previous models. Unfortunately, the accuracy of these models is lower than that of the previous models.

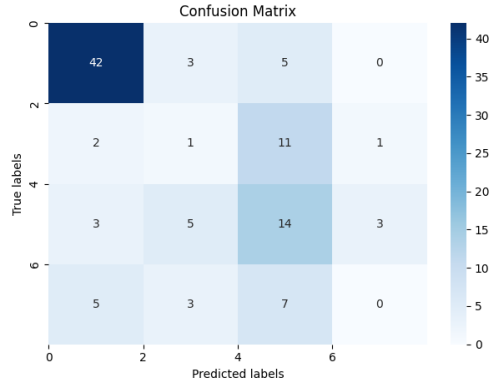


Figure 38: Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data without cognition and with PCA

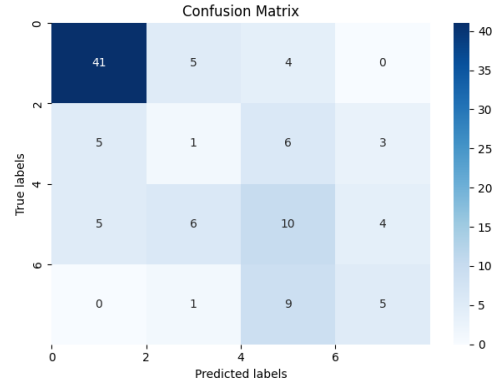


Figure 39: Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data with PCA

To explore the impact of dimension reduction via PCA, two additional tests were conducted without this reduction. Unlike the previous pair of tests that incorporated dimension reduction, this time, the same two tests were performed without it. As a refresher, the first test used all available data except for cognition, while the second test used all available data.

Across 5 runs, the first test achieved a mean accuracy of 0.5429 with a standard deviation of 0.0571. Meanwhile, the second test displayed a mean accuracy of 0.5429 with a standard deviation of 0.0933. Figures 40 and 41 showcase the combined confusion matrices for the first and second tests, respectively. The two figures suggest that this neural network architecture exhibits a reduced inclination toward overfitting than the models made with the SVM or the MLP with only one network. However, it comes at the cost of lower accuracy compared to other models.

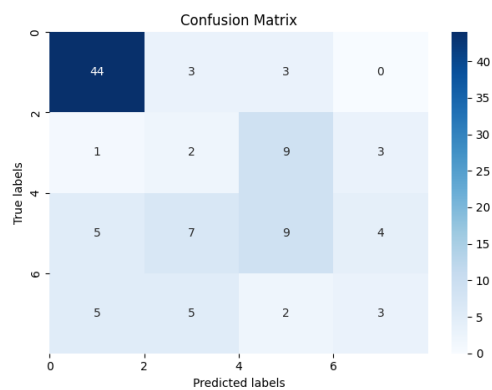


Figure 40: Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data without cognition and without PCA

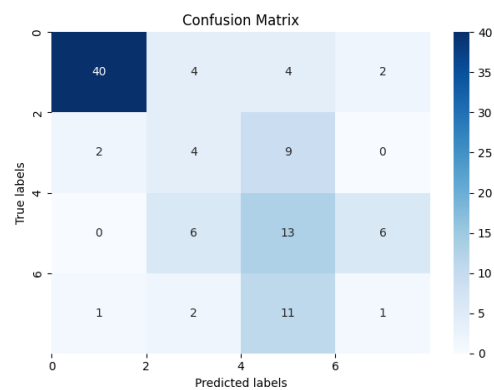


Figure 41: Combined confusion matrix of 5 runs to predict the onset of dementia. Prediction made by an MLP with the z-score normalization and all the data without a PCA

Discussion

8 Visualization

For visualization purposes, the two algorithms will be analyzed at the same time, since the analysis of the different graphs is very similar even though the algorithm used is different.

In Figure 18 and 22, some of the point are in the wrong cluster and throughout all the hyperparameters that have been tested. It was consistently observed that certain points were always positioned in the wrong groups. To investigate this further, the IDs associated with these points were examined to determine if there was a consistent pattern across multiple runs. Remarkably, it was consistently the same ID that appeared in the wrong group of label across all runs of the experiment. This finding suggests a persistent misplacement of a particular ID within the clustering results. It is noteworthy that there are slightly fewer misplaced points with the PCA algorithm than with t-SNE. However, all the IDs misplaced in PCA are also misplaced in t-SNE. Given that the t-SNE algorithm identified additional patients, the analysis of outliers will encompass all outliers detected by the t-SNE algorithm.

To understand why there were two non-demented patients positioned in the middle of the demented patients, An analysis on the data were conducted. For data analysis, a comprehensive CSV file was generated, comprising normalized data exclusively for the non-demented patients. Following that, each cell in the data was assessed to identify instances that deviated by more than two standard deviations from the mean of their respective columns. This analysis unveiled that the two individuals positioned among the demented patients showcased notably high values for the SUVR in PET amyloid scans. This observation provides a possible explanation for their placement among the demented patients, as the high SUVR values indicate abnormal levels of amyloid in their brains.

Subsequently, a similar approach was employed for the other patients who were classified in the middle of the non-demented group instead of the demented group. However, the results were less apparent in this case. For deeper insights,

the diagnoses of these patients were examined. Interestingly, out of the 10 patients who were misplaced, 8 were diagnosed with dementia, but a different form than AD. This finding suggests that the misplacement of these patients can be attributed to their specific type of dementia rather than being non-demented. In the machine learning model for dementia due to AD predictions, those 8 patients were excluded as they did not develop dementia due to AD.

As for the remaining 2 patients, no clear explanation was found. Although they exhibited some deviations from the mean values, they did not show significant differences comparable to the two patients described earlier. These cases require further investigation to determine the factors contributing to their misplacement.

Let's shift our attention to Figure 19 and 23, for this specific dimension and employing the two selected algorithms, it's apparent that no distinct clusters can be identified, except for the one representing non-dementia patients. The labels representing the different year before the onset dementia in the patient population are not grouped together or separated from each other as expected.

A quick word about the 3D plot in Figure 20. It's impossible to distinct a pattern in this 3D plot even with all the different parameters that have been tested. This is the best result achieved and the data is still in an agglomerate sphere with some outliers at the bottom of the figure. However, the 3D graphs created using the PCA algorithm are more easily understood. Figure 24 shows two distinct clusters between demented and non-demented patients. On the other hand, in Figure 25, when attempting to work with the original labels representing the different years before the onset of dementia in the patient population, no additional clusters can be observed apart from the one from the non-dementia patients.

After that the missing value in the cognition test have been replaced. A plot of the data with the cognition test included have been made. However, it result in 2 plot with no clusters that can be distinguish, as you can see in Figure 21 and Figure 26, even with the modified labels to see if a patient will evolve or not toward AD dementia.

Good results are expected for separating patients with dementia from those without, given the clear clusters that were able to be observed on the different graphs. However it is more difficult to draw a conclusion for the prediction of the onset of dementia as no clear clusters could be observed. It's crucial to note that if a visualization technique fails to display a clear cluster, it doesn't necessarily imply that identifying clusters will be impossible later on.

9 Predictive Machine Learning Algorithms

9.1 SVM

The performance evaluation of our predictive models for assessing whether the patient will progress to dementia due to AD showed promising results. Our first classifier uses SVM with data from the MRI and both PET scans. This model showed an accuracy of 0.919 out of 10 runs. The subsequent test, integrating cognition data into the model, showed a notable improvement, increasing the accuracy to 0.9524 out of 10 runs. These results are slightly better than those found in the literature review in section 3 of the state of the art. which obtained an accuracy of 94.5% [70], slightly lower than the 95.2% obtained in this study.

It's important to note that the dataset has balanced classes, ensuring equal representation of patients who progressed to dementia and those who did not. The robust performance of the models is reflected in the ROC curves, with an AUC of 1.00 for the second test, illustrating the ability of the models to discriminate between the two classes, further validating their reliability in predictive accuracy.

However, this is not the main focus of our work. When trying to predict the onset of dementia, the results deteriorate. The tests here were conducted using data that underwent dimension reduction via PCA. As mentioned in the result, the scores drop to a mean accuracy of 0.5714 from the 10 runs with min-max normalization and all data. It improves slightly with z-score normalization instead of min-max. This method gives a mean accuracy of 0.6381. Removing the cognition test from the data was attempted to observe any potential improvement in the result. However, the mean accuracy declined to 0.6238.

Since the SVM is an algorithm that can work with high dimensionality, a test was made to see if not reducing the data would improve the results. The result is slightly lower than with dimension reduction, an average precision of 0.6095 is achieved with all data except the cognition data, while the second test with all data achieves a precision of 0.6143.

Assessing the utility of the cognition test in predicting the onset of dementia in AD is challenging. The data shows a decrease of 1.5% when implementing a dimension reduction method. However, when the cognition test is added to the data without dimension reduction, the results increase by 1.5%. Further investigation is necessary to evaluate the usefulness of the cognition test in predicting the onset of dementia.

Incorporating PCA into all models yielded superior results compared to models without PCA. This improvement could potentially be attributed to the high dimensionality of the MRI data. The dataset initially comprises 464 features, which, through PCA, can be condensed to approximately 50 features while retaining 90% of the original variance, as demonstrated in Figure 16. In contrast, other data sources typically consist of fewer than 100 features. The models may thus be sensitive to the noise inherent in MRI data. However, this proposition remains a hypothesis, necessitating further investigation for confirmation.

When examining the combined confusion matrices in figures 28, 29, 30, 32 and 33. It can be observed that the models tend to predict that the patient will not develop dementia or will develop it in 4 years, with a clear tendency to almost never predict dementia after 6 years. This may be due to the imbalance of the different classes. Approximately half of the patients in the dataset, specifically 51 patients, are not expected to develop dementia. Meanwhile, 17 patients are known to develop dementia after 2 years, 25 after 4 years and 12 after 6 years. On the precision and recall graph, it is evident that patients who will not develop dementia have a high AUC, indicating accurate predictions by the models. However, the AUC for other patients is not satisfactory.

In contrast to the study [72] outlined in the literature review that achieved a mean square error of 1.5 years, our method appears to be less precise. This discrepancy could primarily stem from the model's performance. However, another potential factor could be the classification method, given that our data is structured with time steps occurring every 2 years, unlike the regression approach employed in their study.

9.2 MLP

In the first part of this section the attention will be on the discussion of the results of the first MLP architecture with only one network of neurons. In the second part, the discussion will focus on the results of the second architecture using the 5 neural networks. To evaluate the model's quality, 5 runs were conducted instead of the 10 runs performed for the SVM. This adjustment was necessary as training the MLP is significantly more time-consuming. For the first MLP architecture, k-fold have been implemented and 5 different hyperparameters for the model have been tested in the same execution. The generation of the combined confusion matrix requires 30 minutes to run. The second architecture encounters issues with k-fold implementation, thus only one hyperparameter is tested per run, taking approximately 10 minutes to produce results.

In the first architecture, four tests were conducted to analyze the impact of PCA and cognition data on dementia onset prediction. The results were as follows:

- Test 1 with PCA, excluding cognition data: precision of 0.6286
- Test 2 with PCA and all data: precision of 0.6095
- Test 3 without PCA, excluding cognition data: precision of 0.619
- Test 4 without PCA and all data: precision of 0.6667

The MLP algorithm performs better with this architecture when PCA is not used. However, it is difficult to draw a conclusion about the impact of cognition data on the prediction. The model performs worse with cognition data when using PCA, but improves by 5% without PCA. It is important to note that this conclusion is based on the available data and model used here, but with other parameters, this conclusion may not hold.

Looking at figures 34, 35, 36 and 37, it can be seen that the models have a clear tendency to predict that the patient will not reach dementia or will reach dementia after 4 years. Furthermore, the models never predict that the patient will progress to dementia after 6 years. As for the SVM, the explanation could lie in the class imbalance of our data.

For the second architecture, the four tests, with or without PCA and with or without the cognition data, achieved a precision of around 55%. This is lower than all the previous models discussed in this thesis. However, upon examining Figures 39, 40, 41, and 38, it appears that this architecture is less susceptible to bias caused by class imbalance. Further investigation is necessary to confirm this. It is uncertain whether improving certain hyperparameters to increase model accuracy will prevent bias caused by class imbalance.

In comparison to our MLP models, the method used in the study of Murat Bilgel and Bruno M. Jedynak [72], appears to demonstrate better precision. This discrepancy might stem primarily from the model's performance. However, an alternative factor could be the classification methodology, given that our data is structured with time steps occurring every 2 years, unlike the regression approach used in the referenced study.

10 limitation of the study

This section presents the limitations of this thesis. The first limitation of this study is the small sample size, with only 105 participants remaining in the dataset after excluding certain patients. The study's statistical power and generalizability could be impacted by the limited sample size.

Furthermore, the dataset is imbalanced, resulting in bias towards the majority class in the classifier. The classifier's performance is often affected by this imbalance, particularly when it comes to accurately predicting outcomes for the minority class.

Another notable issue is the tendency of the models to overfit. The superior performance observed on the training data compared to the validation or test sets suggests potential overfitting, raising concerns about the ability of the model to generalize to new data.

Although the chosen algorithms are in line with the current state of the art, it's worth noting that alternative algorithms may give better results. Or, adapting the existing models with different hyperparameters than those tested in this thesis could improve the predictive capabilities of the models and improve the overall performance. change to improve the result.

11 Perspectives

To enhance the results, this section will introduce various potential approaches. The first one is to try to solve the limitation of the number of patients, this can be achieved through 2 methods. The first one is to add the data from the ADNI database. The second solution is to implement data augmentation techniques on the actual data. This two method will increase the number of patient and hopefully improve the ability of the model to generalize to new data.

With regard to the second area of improvement, which could be to address the data imbalance, Three promising methods emerge. First, the use of data augmentation techniques on the minority class could be beneficial by introducing more learning examples for them. Second, adjusting the cost of misclassification specifically for the minority classes could be explored as another viable option. The last approach could be to partition the classifiers to improve the predictions. For example, once the model has accurately identified patients at risk of developing dementia, a subsequent classifier predicts the specific timing of this progression. This reduces class imbalance, as the class of non-dementia patients is excluded from the training of the second classifier.

Another possibility for improvement could be to use PET tau and therefore to use the TAU study instead of the biomarker study. However, for the moment, few patients in the TAU study have reached dementia or have had a sufficient number of timestep to be considered as not having dementia.

Finally, experimenting with alternative machine learning algorithms or delving into different hyperparameters for the existing algorithms remains an option worth exploring.

Conclusion

The rising incidence of AD is a major challenge for healthcare systems worldwide. This neurodegenerative disease progressively impairs cognitive abilities such as memory, language, spatial awareness and problem-solving skills, ultimately leading to dementia. As the leading cause of dementia, AD becomes more likely with age, a concern that is exacerbated by the increasing global life expectancy. Predicting and understanding the progression of AD is critical to developing early interventions and tailored care plans, essential steps in alleviating its profound impact on individuals and healthcare systems worldwide.

The literature review provided an introduction to AD in order to facilitate the understanding of the data used in the thesis. This was followed by a brief exploration of the algorithms used for visualization and prediction. Finally, the review concludes with an overview of related research efforts centred on the prediction of dementia.

In this thesis a part of the originally planned database was created. After some time, the work shifted towards machine learning prediction with the data collected in the first part. The aim was to predict the onset of dementia caused by AD. The choice was made to implement an SVM and two different architectures of MLP to achieve this goal.

The result achieves a precision of around 95% in predicting whether the patient will develop dementia or not. However, when trying to predict the onset of dementia, the score drops to around 60% for the majority of the models tested. Furthermore, the prediction seems to be greatly influenced by the class imbalance. The only models that don't seem to be affected as much by class imbalance only reach a precision of around 55%.

Finally, some suggestions for further development have been made. Firstly, considering the addition of patients from the ADNI dataset to the database or the use of data augmentation techniques may be beneficial. Data augmentation may also be a viable strategy to address class imbalance within datasets. Another way to address class imbalance is by adjusting the cost of misclassification specifically for the minority classes. Finally, additional algorithms or new hyperparameters for the already implemented algorithms can be tested.

Bibliography

- [1] NIH National Institute on Aging. What are the signs of alzheimer’s disease. <https://www.nia.nih.gov/health/what-are-signs-alzheimers-disease>, 2023.
- [2] A. Kumar, J. Sidhu, A. Goyal, and J.W. Tsao. Alzheimer disease. In *StatPearls*. StatPearls Publishing, Treasure Island, FL, 2022.
- [3] Alzheimer’s association. What is alzheimer’s disease ? <https://www.alz.org/alzheimers-dementia/what-is-alzheimers>. Accessed on April 29, 2023.
- [4] World Health Organization. Dementia cases set to triple by 2050 but still largely ignored. <https://www.who.int/news/item/11-04-2012-dementia-cases-set-to-triple-by-2050-but-still-largely-ignored>, 2012. Accessed on April 29, 2023.
- [5] OpenAI. Chatgpt. AI Language Model, 2021.
- [6] DeepL GmbH. Deepl. AI Language Translation Service, 2021.
- [7] Alzheimer’s association. 2023 alzheimer’s disease facts and figures. <https://www.alz.org/alzheimers-dementia/what-is-alzheimers>.
- [8] Richard Armstrong. Risk factors for alzheimer’s disease. *Folia neuropathologica*, 57(2):87–105, 2019.
- [9] Lewis O. Killin, John M. Starr, Ivy J. Shiue, and Tom C. Russ. Environmental risk factors for dementia: A systematic review. *BMC Geriatrics*, 16(1):175, Oct 2016.
- [10] Deepti Bhardwaj, Chiranjib Mitra, Chandrakala A Narasimhulu, Ahmed Riad, Manu Doomra, and Sampath Parthasarathy. Alzheimer’s disease-current status and future directions. *Journal of medicinal food*, 20(12):1141–1151, 2017.

- [11] Maileen Gloriane Ulep, Simrit Kaur Saraon, and Samantha McLea. Alzheimer disease. *The Journal for Nurse Practitioners*, 14(3):129–135, 2018. DEMENTIA.
- [12] Caitlin Beddows. Neural basis of alzheimer’s disease. <https://scitechconnect.elsevier.com/neural-basis-alzheimers-disease/>, 2016.
- [13] Arnold Loof and Liliane Schoofs. Alzheimer’s disease: Is a dysfunctional mevalonate biosynthetic pathway the master-inducer of deleterious changes in cell physiology? *OBM Neurobiology*, 3:1–1, 10 2019.
- [14] David Eisenberg and Mathias Jucker. The amyloid state of proteins in human diseases. *Cell*, 148(6):1188–1203, 2012.
- [15] Liana G. Apostolova. Alzheimer disease. *Continuum (Minneapolis, Minn.)*, 22(2 Dementia):419–434, Apr 2016.
- [16] G.K. Gouras, T.T. Olsson, and O. Hansson. -amyloid peptides and amyloid plaques in alzheimer’s disease. *Neurotherapeutics*, 12:3–11, 2015.
- [17] J. P. Bonton. The role of neurofibrillary tangles in alzheimer disease. *Acta Neurologica Belgica*, 98:165–174, 1998.
- [18] W.K. Self and D.M. Holtzman. Emerging diagnostics and therapeutics for alzheimer disease. *Nature Medicine*, 29:2187–2199, 2023.
- [19] NIH National Institute on Aging. What happens to the brain in alzheimer’s disease? <https://www.nia.nih.gov/health/what-happens-brain-alzheimers-disease>, Accessed 2023.
- [20] Flavius Raslau, I Mark, Andrew Klein, J Ulmer, V Mathews, and L Mark. Memory part 2: The role of the medial temporal lobe. *AJNR. American journal of neuroradiology*, 36, 11 2014.
- [21] Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics*, 69:89–95, 2001.
- [22] Smailagic N Roqué i Figuls M Ciapponi A Sanchez [U+2010] Perez E Giannakou A Pedraza OL Bonfill Cosp X Arevalo [U+2010] Rodriguez, I and S Cullum. Mini [U+2010] mental state examination (mmse) for the detection of alzheimer’s disease and other dementias in people with mild cognitive impairment (mci). *Cochrane Database of Systematic Reviews*, (3), 2015.

- [23] Tamara G. Fong, Michael A. Fearing, Richard N. Jones, Peilin Shi, Edward R. Marcantonio, James L. Rudolph, Frances M. Yang, Dan K. Kiely, and Sharon K. Inouye. Telephone interview for cognitive status: Creating a crosswalk with the mini-mental state examination. *Alzheimer's Dementia*, 5(6):492–497, 2009.
- [24] Marc Teichmann, Stéphane Epelbaum, Dalila Samri, Marcel Levy Nogueira, Agnès Michon, Harald Hampel, Foudil Lamari, and Bruno Dubois. Free and cued selective reminding test – accuracy for the differential diagnosis of alzheimer's and neurodegenerative diseases: A large-scale biomarker-characterized monocenter cohort study (clinad). *Alzheimer's Dementia*, 13(8):913–923, 2017.
- [25] Kevin Duff, Danielle Tometich, and Kathryn Dennett. The modified telephone interview for cognitive status is more predictive of memory abilities than the mini-mental state examination. *Journal of Geriatric Psychiatry and Neurology*, 28(3):193–197, 2015.
- [26] Kesheng Wang, Chun Xu, Amanda Smith, Danqing Xiao, R. Osvaldo Navia, Yongke Lu, Changchun Xie, and Ubolrat Piamjariyakul. Genome-wide association study identified insc gene associated with trail making test part a and alzheimer's disease related cognitive phenotypes. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 111:110393, 2021.
- [27] Betsy White Williams, Wendy Mack, and Victor W. Henderson. Boston naming test in alzheimer's disease. *Neuropsychologia*, 27(8):1073–1079, 1989.
- [28] S.J. Duff Canning, L. Leach, D. Stuss, L. Ngo, and S. E. Black. Diagnostic utility of abbreviated fluency measures in alzheimer disease and vascular dementia. *Neurology*, 62(4):556–562, 2004.
- [29] Gisele P. Wolf-Klein, Felix A. Silverstone, Arnold P. Levy, Meryl S. Brod, and Joseph Breuer. Screening for alzheimer's disease by clock drawing. *Journal of the American Geriatrics Society*, 37(8):743–746, August 1989.
- [30] K. A. Welsh, N. Butters, R. C. Mohs, D. Beekly, S. Edland, G. Fillenbaum, and A. Heyman. The consortium to establish a registry for alzheimer's disease (cerad). part v. a normative study of the neuropsychological battery. *Neurology*, 44(4):609–609, 1994.
- [31] Sarah Tomaszewski Farias, Dan Mungas, Danielle J. Harvey, Andy Simmons, Bruce R. Reed, and Charles Decarli. The measurement of everyday cognition: development and validation of a short form of the everyday cognition scales. *Alzheimer's & Dementia*, 7(6):593–601, Nov 2011.

- [32] David Tulskey, Nancy Chiaravalloti, Barton Palmer, and Gordon Chelune. *The Wechsler Memory Scale, Third Edition*, pages 93–139. 12 2003.
- [33] Sarah E. MacPherson, Martha S. Turner, Marco Bozzali, Lisa Cipolotti, and Tim Shallice. The doors and people test: The effect of frontal lobe lesions on recall and recognition memory performance. *Neuropsychology*, 30(3):332–337, March 2016.
- [34] MD Preston, David C. Magnetic resonance imaging (mri) of the brain and spine: Basics. <https://case.edu/med/neurology/NR/MRI2023>. Revised 07/04/16. Copyrighted 2006, David C. Preston, MD.
- [35] National Institute of Biomedical Imaging and Bioengineering (NIBIB). Magnetic resonance imaging (mri). <https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri>, Accessed 2023.
- [36] Laboratory for Computational Neuroimaging at the Athinoula A. Martinos Center for Biomedical Imaging. FreeSurfer. <https://surfer.nmr.mgh.harvard.edu/>, Accessed 2023.
- [37] Mariano Cabezas, Arnau Oliver, Xavier Lladó, Jordi Freixenet, and Meritxell Bach Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Computer Methods and Programs in Biomedicine*, 104(3):e158–e177, 2011.
- [38] Andy Jahn. Freesurfer roi analysis, 2023.
- [39] Gary H. Glover. Overview of functional magnetic resonance imaging. *Neurosurgical Clinics of North America*, 22(2):133–139, vii, Apr 2011.
- [40] R. et al. Goebel. Brainvoyager (brain innovation, maastricht, the netherlands), 2006.
- [41] Maja A.A. Binnewijzend, Menno M. Schoonheim, Ernesto Sanz-Arigita, Alle Meije Wink, Wiesje M. van der Flier, Nelleke Tolboom, Sofie M. Adriaanse, Jessica S. Damoiseaux, Philip Scheltens, Bart N.M. van Berckel, and Frederik Barkhof. Resting-state fmri changes in alzheimer’s disease and mild cognitive impairment. *Neurobiology of Aging*, 33(9):2018–2028, 2012.
- [42] Lauren J. O’Donnell and Carl-Fredrik Westin. An introduction to diffusion tensor image analysis. *Neurosurgical Clinics of North America*, 22(2):185–196, viii, Apr 2011.

- [43] Elisa Scola, Marco Bozzali, Federica Agosta, Giuseppe Magnani, Marco Franceschi, Maria Pia Sormani, and Andrea Falini. A diffusion tensor mri study of patients with mci and ad with a 2-year clinical follow-up. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(7):798–805, 2010.
- [44] K Lameka, MD Farwell, and M Ichise. Positron emission tomography. *Handbook of Clinical Neurology*, 135:209–227, 2016.
- [45] Charles Marcus, Esther Mena, and Rathan M. Subramaniam. Brain pet in the diagnosis of alzheimer’s disease. *Clinical Nuclear Medicine*, 39(10):e413–e422; quiz e423–6, Oct 2014.
- [46] Bernard J Hanseeuw, Rebecca A Betensky, Heidi IL Jacobs, Aaron P Schultz, Jorge Sepulcre, J Alex Becker, Danielle M Cosio, Michelle Farrell, Yakeel T Quiroz, Elizabeth C Mormino, Rachel F Buckley, Kathryn V Papp, Rebecca A Amariglio, Ilse Dewachter, Adrian Ivanoiu, Willem Huijbers, Trey Hedden, Gad A Marshall, Jasmeer P Chhatwal, Dorene M Rentz, Reisa A Sperling, and Keith Johnson. Association of amyloid and tau with cognition in preclinical alzheimer disease: A longitudinal study. *JAMA Neurol*, 76(8):915–924, 2019.
- [47] Michael Schöll, Samuel N Lockhart, Daniel R Schonhaut, James P O’Neil, Mustafa Janabi, Rik Ossenkoppele, Suzanne L Baker, Jacob W Vogel, Jamie Faria, Henry D Schwimmer, et al. Pet imaging of tau deposition in the aging human brain. *Neuron*, 89(5):971–982, 2016.
- [48] Heiko Braak and Eva Braak. Neuropathological staging of alzheimer-related changes. *Acta Neuropathol*, 82:239–259, 1991.
- [49] Prashanthi Vemuri, Val J. Lowe, David S. Knopman, Matthew L. Senjem, Bradley J. Kemp, Christopher G. Schwarz, Scott A. Przybelski, Mary M. Machulda, Ronald C. Petersen, and Clifford R. Jack. Tau-pet uptake: Regional variation in average suvr and impact of amyloid deposition. *Alzheimer’s Dementia: Diagnosis, Assessment Disease Monitoring*, 6:21–30, 2017.
- [50] Kjell Erlandsson, Irène Buvat, P. Hendrik Pretorius, Benjamin A Thomas, and Brian F Hutton. A review of partial volume correction techniques for emission tomography and their applications in neurology, cardiology and oncology. *Physics in Medicine & Biology*, 57(21):R119, October 2012.
- [51] Yi Su, Tyler M. Blazey, Abraham Z. Snyder, Marcus E. Raichle, Daniel S. Marcus, Beau M. Ances, Randall J. Bateman, Nigel J. Cairns, Patricia Aldea, Lisa Cash, Jon J. Christensen, Karl Friedrichsen, Russ C. Hornbeck, Angela M. Farrar, Christopher J. Owen, Richard Mayeux, Adam M. Brickman, William

- Klunk, Julie C. Price, Paul M. Thompson, Bernadino Ghetti, Andrew J. Saykin, Reisa A. Sperling, Keith A. Johnson, Peter R. Schofield, Virginia Buckles, John C. Morris, and Tammie L.S. Benzinger. Partial volume correction in quantitative amyloid imaging. *NeuroImage*, 107:55–64, 2015.
- [52] Oskar Hansson, Sylvain Lehmann, Markus Otto, et al. Advantages and disadvantages of the use of the csf amyloid (a) 42/40 ratio in the diagnosis of alzheimer’s disease. *Alzheimer’s Research & Therapy*, 11:34, 2019.
- [53] David S. Knopman, Helene Amieva, Ronald C. Petersen, et al. Alzheimer disease. *Nature Reviews Disease Primers*, 7, 2021.
- [54] Clive Ballard, Serge Gauthier, Anne Corbett, Carol Brayne, Dag Aarsland, and Emma Jones. Alzheimer’s disease. *The Lancet*, 377(9770):1019–1031, Mar 2011.
- [55] Ming Shi, Fei Chu, Fang Zhu, and Jing Zhu. Impact of anti-amyloid- monoclonal antibodies on the pathology and clinical profile of alzheimer’s disease: A focus on aducanumab and lecanemab. *Frontiers in Aging Neuroscience*, 14, 2022.
- [56] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.
- [57] John A. Lee and Michel Verleysen. *Nonlinear Dimensionality Reduction*, volume 1. Springer, New York, 2007.
- [58] Principal component analysis (pca) and singular value decomposition (svd).
- [59] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.
- [60] Abdulrahman Abbas, Belal Al-Khateeb, and Mazin Mohammed. An extensive review of state-of-the-art transfer learning techniques used in medical imaging: Open issues and challenges. *Journal of Intelligent Systems*, 31:1085–1111, 09 2022.
- [61] Daniel Berrar. Cross-validation. Data Science Laboratory, Tokyo Institute of Technology, 2-12-1-S3-70 Ookayama, Meguro-ku, Tokyo 152-8550, Japan, 2019.
- [62] William Noble. What is a support vector machine? *Nature Biotechnology*, 24:1565–1567, 2006.

- [63] MathWorks. Support vector machine. <https://se.mathworks.com/discovery/support-vector-machine.html>, Accessed 2023.
- [64] S. Suthaharan. Support vector machine. In *Machine Learning Models and Algorithms for Big Data Classification*, volume 36 of *Integrated Series in Information Systems*. Springer, Boston, MA, 2016.
- [65] Alessia Mammone, Marco Turchi, and Nello Cristianini. Support vector machines. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):283–289, November/December 2009.
- [66] Marius-Constantin Popescu, Valentina E. Balas, Liliana Perescu-Popescu, et al. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7):579–588, 2009.
- [67] Haitham Afan, Ahmedbahaaaldin Ibrahim Ahmed Osman, Yusuf Essam, Al-Mahfoodh Najah, Yuk Huang, Ozgur Kisi, Mohsen Sherif, Ahmed Sefelnasr, Kwok Chau, and Ahmed El-Shafie. Modeling the fluctuations of groundwater level by employing ensemble deep learning techniques. *Engineering Applications of Computational Fluid Mechanics*, 15:1420–1439, 09 2021.
- [68] Perceptron : qu'est-ce que c'est et à quoi ça sert ? <https://datascientest.com/perceptron>, Accessed 2023.
- [69] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. Activation functions in neural networks. *Towards Data Science*, 6(12):310–316, 2017.
- [70] B. Magnin, L. Mesrob, S. Kinkingnéhun, et al. Support vector machine-based classification of alzheimer's disease from whole-brain anatomical mri. *Neuroradiology*, 51:73–83, 2009.
- [71] S. Grueso and R. Viejo-Sobera. Machine learning methods for predicting progression from mild cognitive impairment to alzheimer's disease dementia: a systematic review. *Alzheimers Research & Therapy*, 13:162, 2021.
- [72] Murat Bilgel, Bruno M. Jedynak, and Alzheimer's Disease Neuroimaging Initiative. Predicting time to dementia using a quantitative template of disease progression. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 11(1):205–215, 2019.
- [73] Sriramakrishnan Padmanaban, Kalaiselvi Thiruvankadam, Padmapriya T., M. Thirumalaiselvi, and Ramkumar Sivasakthivel. A role of medical imaging techniques in human brain tumor treatment. 8:565–568, 01 2020.

- [74] Hwamee Oh. Head-to-head harmonization of tau pet tracers – the head study. Carney Institute for Brain Science, Brown University.
- [75] Wes McKinney et al. pandas, 2020.
- [76] *Data Preprocessing*, chapter 2, pages 16–50. John Wiley Sons, Ltd, 2014.
- [77] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [78] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [79] FreeSurfer Development Team. PetSurfer. <https://surfer.nmr.mgh.harvard.edu/fswiki/PetSurfer>, Accessed 2023.

Appendix

12 Dataset creation

12.1 MRI Processing

As explained earlier in this thesis, freesurfer uses the MRI data of the patient and processes it through a series of computational steps. The software generates various text files as output, depending on the specific analyses requested by the researcher. In this section we will present how the output text files of freesurfer have been turned into several csv files.

In our work, we had to obtain the patients' files and standardize the folder organization to maintain consistency. Since the research team used different versions of Freesurfer, we had to segregate the data into two groups, each corresponding to the respective version used. Additionally, it was crucial for us to keep track of the version information for accurate analysis and interpretation of the results. By organizing the data in this manner, we ensured a unified approach to handling the output files from Freesurfer while accommodating the variations arising from different software versions.

The text files outputted by Freesurfer are organized as shown in Figure 42. We have identified four distinct types of structures present in the Freesurfer files. However, most of the files adhere to the structure shown below. We have successfully implemented the necessary code to extract the data from each structure, and all the collected data has been efficiently stored in CSV format for further analysis and processing.

As said before the figure 42 provides a comprehensive overview of the structure of the most common text file generated by Freesurfer. It is observed that the text file consists of distinct sections.

The first part (highlighted in red) contains information about the Freesurfer configuration.

The second part (highlighted in blue) provides general information about the

patient's brain.

Following that (highlighted in green), this section provides a dictionary that present the column of the table that will follow. It contains the name of the column the abbreviation used in the table as well as the units of the column.

Lastly (highlighted in brown), the table presents various measurements such as brain thickness, surface area, gray matter volume, etc. these measurements are specific to different brain regions categorized according to several atlases (An atltals is a specific subdivision of the brain). In our study, we considered the following atlases, each of which has a file for both the right and left hemisphere of the brain

- `apparc_pial`
- `BA_exvivo`
- `BA_exvivo_thresh`
- `desikan_killiany`
- `desikan_killiany_tourville`
- `destrieux`

```
# Table of FreeSurfer cortical parcellation anatomical statistics
#
# CreationTime 2019/02/02-17:47:11-GMT
# generating_program mris_anatomical_stats
# cvs_version $Id: mris_anatomical_stats.c,v 1.79 2016/03/14 15:15:34 greve Exp $
# mrisurf.c-cvs_version $Id: mrisurf.c,v 1.781.2.6 2016/12/27 16:47:14 zkaufman Exp $
# cmdline mris_anatomical_stats -th3 -mgz -cortex ../label/lh.cortex.label -f ../stats/lh.aparc.DKTatlas.stats
# -b -a ../label/lh.aparc.DKTatlas.annot -c ../label/aparc.annot.DKTatlas.ctab subject_name lh white
# sysname Linux
# hostname freecrusher3
# machine x86_64
# user user
```

```
# SUBJECTS_DIR /home/user/freesurfer/stable-v6.0.0/subjects
# anatomy_type surface
# subjectname subject_name
# hemi lh
# AnnotationFile ../label/lh.aparc.DKTatlas.annot
# AnnotationFileTimeStamp 2019/02/02 18:46:47
# Measure Cortex, NumVert, Number of Vertices, 128141, unitless
# Measure Cortex, WhiteSurfArea, White Surface Total Area, 87450.2, mm^2
# Measure Cortex, MeanThickness, Mean Thickness, 2.36461, mm
# Measure BrainSeg, BrainSegVol, Brain Segmentation Volume, 1171058.000000, mm^3
# Measure BrainSegNotVent, BrainSegVolNotVent, Brain Segmentation Volume Without Ventricles, 1116212.000000, mm^3
# Measure BrainSegNotVentSurf, BrainSegVolNotVentSurf, Brain Segmentation Volume Without Ventricles from Surf, 1116161.022221, mm^3
# Measure Cortex, CortexVol Total cortical gray matter volume, 461098.292666, mm^3
# Measure SupraTentorial, SupraTentorialVol, Supratentorial volume, 1038024.022221, mm^3
# Measure SupraTentorialNotVent, SupraTentorialVolNotVent, Supratentorial volume, 989350.022221, mm^3
# Measure EstimatedTotalIntraCranialVol, eTIV, Estimated Total Intracranial Volume, 1476460.114226, mm^3
```

```
# NTableCols 10
# TableCol 1 ColHeader StructName
# TableCol 1 FieldName Structure Name
# TableCol 1 Units NA
# TableCol 2 ColHeader NumVert
# TableCol 2 FieldName Number of Vertices
# TableCol 2 Units unitless
# TableCol 3 ColHeader SurfArea
# TableCol 3 FieldName Surface Area
# TableCol 3 Units mm^2
# TableCol 4 ColHeader GrayVol
# TableCol 4 FieldName Gray Matter Volume
# TableCol 4 Units mm^3
# TableCol 5 ColHeader ThickAvg
# TableCol 5 FieldName Average Thickness
# TableCol 5 Units mm
# TableCol 6 ColHeader ThickStd
# TableCol 6 FieldName Thickness StdDev
# TableCol 6 Units mm
# TableCol 7 ColHeader MeanCurv
# TableCol 7 FieldName Integrated Rectified Mean Curvature
# TableCol 7 Units mm^-1
# TableCol 8 ColHeader GausCurv
# TableCol 8 FieldName Integrated Rectified Gaussian Curvature
# TableCol 8 Units mm^-2
# TableCol 9 ColHeader FoldInd
# TableCol 9 FieldName Folding Index
# TableCol 9 Units unitless
# TableCol 10 ColHeader CurvInd
# TableCol 10 FieldName Intrinsic Curvature Index
# TableCol 10 Units unitless
```

#	ColHeaders	StructName	NumVert	SurfArea	GrayVol	ThickAvg	ThickStd	MeanCurv	GausCurv	FoldInd	CurvInd	
		caudalanteriorcingulate		1324	909	2245	2.651	0.505	0.139	0.020	26	1.0
		caudalmiddlefrontal		3928	2723	6432	2.279	0.483	0.128	0.025	41	4.0
		cuneus		3479	2235	4639	1.972	0.510	0.133	0.031	42	4.3
		entorhinal		650	519	2225	3.242	0.742	0.140	0.033	8	0.9
		fusiform		4292	2949	9299	2.757	0.582	0.137	0.030	65	5.4
		inferiorparietal		7133	4938	12423	2.336	0.524	0.130	0.028	92	7.8
		inferiortemporal		5098	3718	11976	2.675	0.652	0.152	0.037	98	8.1
		isthmuscingulate		1716	1181	3011	2.280	1.004	0.133	0.028	29	1.7
		lateraloccipital		8430	5624	13802	2.171	0.582	0.148	0.034	129	12.0
		lateralorbitofrontal		3837	2692	8121	2.568	0.667	0.144	0.043	58	6.7
		lingual		4467	3018	6724	2.063	0.607	0.139	0.032	62	5.8
		medialorbitofrontal		1949	1359	3765	2.305	0.674	0.133	0.035	34	2.7
		middletemporal		5437	3851	10782	2.512	0.481	0.136	0.030	86	7.2
		parahippocampal		1088	769	2856	3.250	0.850	0.126	0.031	12	1.3
		paracentral		2954	1802	4402	2.350	0.510	0.103	0.021	31	2.6
		parsopercularis		2306	1597	4030	2.358	0.466	0.125	0.028	26	2.5
		parsorbitalis		895	621	1843	2.548	0.529	0.138	0.028	10	1.0
		parstriangularis		2361	1646	4126	2.285	0.557	0.137	0.033	29	3.1
		pericalcarine		2009	1324	1865	1.569	0.422	0.146	0.039	27	3.3
		postcentral		7913	5123	11229	2.079	0.624	0.117	0.023	81	7.8
		posteriorcingulate		1735	1160	2976	2.565	0.710	0.131	0.031	24	2.2
		precentral		8678	5631	13631	2.346	0.525	0.121	0.025	87	8.6
		precuneus		5768	3988	10194	2.370	0.583	0.130	0.028	69	6.2
		rostralanteriorcingulate		1558	1028	2554	2.391	0.813	0.134	0.030	26	1.8
		rostralmiddlefrontal		6191	4449	10783	2.076	0.586	0.154	0.042	109	10.8
		superiorfrontal		10884	7391	20856	2.496	0.561	0.131	0.029	131	12.6
		superiorparietal		7160	4729	10705	2.110	0.510	0.116	0.025	78	7.0
		superiortemporal		6621	4792	15047	2.680	0.592	0.116	0.022	81	6.5
		supramarginal		4664	3177	8205	2.453	0.527	0.125	0.025	50	4.6
		transversetemporal		789	489	1330	2.330	0.481	0.111	0.024	8	0.8
		insula		2827	2020	6388	2.992	0.710	0.132	0.039	37	4.8

Figure 42: Illustration of a Freesurfer text file of a patient's left hemisphere (lh.aparc.DKTatlas)

The script to extract data from the text files of Freesurfer has been developed to support both version 6 and version 7 of the software. As a result, the output from these files has been organized into multiple CSV files.

The "General CSV" encompasses essential information about the patient's brain (information found in the blue box of Figure 42).

Additionally, specific CSV files have been created for distinct parts of the brain, such as white matter, subcortical regions, hippocampus, and amygdala.

Moreover, the different possible brain atlases, including apparc pial, BA exvivo, BA exvivo thresh, desikan killiany, desikan killiany tourville, and destrieux, have each been stored in separate CSV files.

When Freesurfer separates the right and left parts of the brain into multiple text files, we chose to merge the data from the same atlas or brain region into a single CSV file. This avoids unnecessary duplication and simplifies data management.

This comprehensive organization ensures a systematic and easily accessible approach to handle the extracted data, facilitating further analysis and interpretation.

12.2 PET scan Processing

In the version 6 of Freesurfer, the folder organization revolves around different types of PET-scan data and their respective corrections. Each patient's folder can contain the following subfolders :

- **amy_gtmpvc_fwhm6.output**: This folder contains the results of the amyloid PET-scan with Voxel-based partial volume correction (PVC) applied.
- **amy_nopvc.output**: Similar to the previous folder, this one contains the results of the amyloid PET-scan without PVC correction.
- **fdg_gtmpvc_fwhm6.output**: This folder contains the results of the FDG PET-scan (fluorodeoxyglucose) with PVC applied.
- **fdg_nopvc.output**: Similar to the previous folder, this one contains the results of the FDG PET-scan without PVC correction.

These folders store the PET-scan data, either with or without the PVC correction, as applicable. There is another folder for the PET TAU if it have been done.

The text file organization is depicted in [79]. As shown in the picture 43 below, the third column represents the name of a specific brain region, while the seventh column displays the PVC uptake of the corresponding region of interest relative to the Pons. Notably, in the last line of Figure 43, it is evident that the reference region, Pons, is indeed set to a fixed value of 1. This standardized reference ensures consistency in the PVC uptake measurements for accurate comparisons across different brain regions.

For the output files, we opted to maintain the same organization, but we made changes to the names due to their length. The names have been shortened to AMY_nopvc, AMY_pvc, FDG_nopvc, and FDG_pvc. Additionally, we transferred the data from the text files into CSV files. This transition to CSV format allows for easier data handling.

1	2	Left-Cerebral-White-Matter	wm	26023	9559.504	0.905	0.0707
2	7	Left-Cerebellum-White-Matter	wm	1571	774.312	0.955	0.1305
3	8	Left-Cerebellum-Cortex	subcort_gm	6419	3862.978	0.322	0.0785
4	10	Left-Thalamus-Proper	subcort_gm	863	482.614	0.563	0.1374
5	11	Left-Caudate	subcort_gm	376	119.962	0.357	0.0695
6	12	Left-Putamen	subcort_gm	576	236.236	0.286	0.0591
7	13	Left-Pallidum	subcort_gm	207	70.282	0.949	0.1091
8	16	Brain-Stem	wm	1431	590.883	0.766	0.1224
9	17	Left-Hippocampus	subcort_gm	515	187.367	0.334	0.0503
10	18	Left-Amygdala	subcort_gm	199	65.893	0.329	0.0360
11	24	CSF	csf	3866	1475.443	0.008	0.0624
12	26	Left-Accumbens-area	subcort_gm	62	14.685	0.037	0.0521
13	28	Left-VentralDC	wm	494	208.016	0.855	0.0889
14	31	Left-choroid-plexus	csf	121	15.279	0.475	0.0647
15	41	Right-Cerebral-White-Matter	wm	26617	9807.964	0.914	0.0749
16	46	Right-Cerebellum-White-Matter	wm	1598	815.704	0.959	0.1533
17	47	Right-Cerebellum-Cortex	subcort_gm	6692	4019.240	0.311	0.0779
18	49	Right-Thalamus-Proper	subcort_gm	809	429.630	0.542	0.1154
19	50	Right-Caudate	subcort_gm	417	151.098	0.317	0.0433
20	51	Right-Putamen	subcort_gm	556	226.167	0.330	0.0694
21	52	Right-Pallidum	subcort_gm	207	67.820	1.018	0.1142
22	53	Right-Hippocampus	subcort_gm	538	203.819	0.447	0.0641
23	54	Right-Amygdala	subcort_gm	221	81.480	0.414	0.0475
24	58	Right-Accumbens-area	subcort_gm	59	11.603	0.125	0.0729
25	60	Right-VentralDC	wm	514	208.850	0.888	0.1179
26	63	Right-choroid-plexus	csf	103	11.158	0.309	0.0693
27	130	Air	head	12061	6747.474	0.177	0.0645
28	165	Skull	head	106860	56221.438	0.226	0.0798
29	172	Vermis	subcort_gm	343	115.453	0.266	0.0485
30	174	Pons	wm	1151	680.127	1.000	0.0800

Figure 43: Illustration of part of a Freesurfer text file for a patient's amyloid PET scan

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain

Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/epl