TÉCNICO LISBOA

Medical neuroimage consolidation from a grid of hospitals for the external validation of predictive models

Rui Pedro Queirós Nóbrega

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Diana Maria Pinto Prata
Prof. Rui Miguel Carrasqueiro Henriques

Examination Committee

President: Prof. António Manuel Ferreira Rito da Silva
Supervisor: Prof. Diana Maria Pinto Prata
Members of the Committee: Prof. Andreas Miroslaus Wichert

November 2021

Acknowledgments

I would like to express my gratitude and appreciation for my supervisor Prof./Dra. Diana Prata and Prof/Dr. Rui Henriques whose guidance, support, and encouragement have been invaluable throughout this dissertation which allowed me to push myself the whole time I worked with them, striving for excellence in every job we put our minds to. I also wish to thank the team at the Institute of Biophysics and Biomedical Engineering (IBEB) who have been a great source of support and the work done with them was marvelous.

I am also thankful to my college friends, without whom this dissertation would not have been such a beautiful and joyful ride during the hard times we faced due to Covid-19. A special thanks to my hometown friends which always supported me in every project I embarked on in my life and this dissertation was an example of that.

Most importantly, I owe my deepest gratitude to my beautiful family who is responsible for this amazing milestone I have achieved by concluding this dissertation. I owe them everything and especially my wonderful and incredible sister who was always there for me, not only in the role of a sister but especially as a friend.

Lastly, I would like to dedicate the most important project in my life to all my friends and family and my beautiful grandparent whose life was set way too short due to Dementia and whose condition inspired the topic of this work.

Abstract

The diagnosis of Alzheimer's disease is only certain with a detailed post-mortem microscopic examination of the brain. Machine learning approaches are increasingly used in the development of predictive models for the early diagnosis of Alzheimer´s disease. The major issues with such models are the lack of interpretability at the clinical end and the lack of generalization of said models due to the heterogeneity of the data sources (instrumentation, monitoring protocol, individual demographics). To tackle these issues, this work proposes a multi-diagnostic, multi-site, clinically interpretable tool using MRI imaging. Furthermore, it presents the steps for the data consolidation where the MRIs are extracted from heterogeneous sources and are anonymized in order to maintain the anonymity of the patients subjected to the study. In addition, the performance of the models is externally validated on data obtained independently according to temporal, geographic, and/or domain differences. The models could not generalize well for the target population as they generalized for the testing partitions of the original data. Out of the three possible class labels, class Control showed the worst results, returning 100% of precision yet significantly low levels of recall. MCI and AD classes returned similar results of precision, 29% and 30% respectively, however, AD had 83% of recall whereas MCI only 43%. The gathered observations confirm the difficulty of performing neuroimaging diagnostics under the different monitoring protocols, medical classifications, and population demographics.

Keywords: Medical Resonance Imaging - Alzheimer´s disease - Mild Cognitive Impairment - Predictive models - External Validation - Data Consolidation.

Resumo

O diagnóstico da doença de Alzheimer só é certo com um exame microscópico post-mortem detalhado do cérebro. As abordagens de Machine Learning são cada vez mais utilizadas no desenvolvimento de modelos preditivos para o diagnóstico precoce da doença de Alzheimer. Os principais problemas com tais modelos são a falta de interpretabilidade no final clínico e a falta de generalização dos referidos modelos devido à heterogeneidade das fontes de dados (instrumentação, protocolos de monitoramento, dados demográficos individuais). Para lidar com essas questões, este trabalho propõe uma ferramenta multi-diagnóstica, multi-lugar, clinicamente interpretável usando imagens de ressonância magnética. Além disso, apresenta as etapas para a consolidação dos dados onde as ressonâncias magnéticas são extraídas de fontes heterogêneas e anonimizadas a fim de manter o anonimato dos pacientes submetidos ao estudo. Além disso, o desempenho dos modelos é validado externamente em dados obtidos de forma independente de acordo com diferenças temporais, geográficas e / ou de domínio. Os modelos não generalizaram bem para a população-alvo tal como generalizaram para os dados originais. Das três classes possíveis, a classe Control apresentou o pior resultado, retornando 100% de precisão, mas apenas 16% de recall. As classes MCI e AD retornaram resultados semelhantes de precisão, 29% e 30% respectivamente, no entanto, a classe AD teve 83% de recall, enquanto o MCI apenas 43%. As observações recolhidas confirmam a dificuldade de realizar diagnósticos de neuroimagens ao abrigo dos diferentes protocolos de monitorização, classificações médicas, e demografia populacional.

Keywords: Imagem por Ressonância Magnética - Doença de Alzheimer - Défice cognitivo leve - Validação externa - Consolidação de dados.

Contents

List of Tables xi

List of Figures xiii

1 Introduction 1
- 1.1 Problem Description . . . 1
- 1.2 Research contributions . . . 2
- 1.3 Outline . . . 2

2 Background 3
- 2.1 Medical Context . . . 3
- 2.2 Data Extraction . . . 4
- 2.3 Data Storage . . . 5
  - 2.3.1 OLTP and OLAP Systems . . . 5
  - 2.3.2 Relational and Multidimensional Databases . . . 5
- 2.4 Data Science Concepts . . . 6
  - 2.4.1 Predictive assessment . . . 6
  - 2.4.2 Predictive modeling . . . 9

3 Related Work 13
- 3.1 Data Consolidation . . . 13
  - 3.1.1 Data security and privacy related issues . . . 14
- 3.2 Predictive model validation . . . 14
  - 3.2.1 Biomarkers Discovery . . . 15
  - 3.2.2 External validation . . . 16
  - 3.2.3 Internal and internal-external validation . . . 17

4 Development 19
- 4.1 Project's Architecture . . . 19
- 4.2 Data Consolidation . . . 20
  - 4.2.1 Data anonymization . . . 20
  - 4.2.2 Data Storage . . . 21
- 4.3 Graphical user interface . . . 22
  - 4.3.1 Input files . . . 22
  - 4.3.2 View Data . . . 22
- 4.4 Models Description . . . 24
  - 4.4.1 Training the models . . . 24
  - 4.4.2 Testing the models . . . 25

4.5 External Validation . . . 26
4.5.1 Measures in training . . . 26
4.5.2 Testing the models on a Portuguese population . . . 28
Target population . . . 28

5 Results and Discussion **31**
5.1 Data anonymization . . . 31
5.2 Database . . . 32
5.3 Statistical Validation . . . 34
5.3.1 Validation using the original (heterogeneous) population . . . 34
Comparing the models with ANOVA . . . 39
5.3.2 Generalization analysis in a Portuguese population . . . 39

6 Conclusion **43**
6.1 Future Work . . . 44

Bibliography **45**

A Data anonymization **49**

B GUI screens **53**

C Database **55**
C.1 Relevant queries . . . 55

D External Validation **57**
D.1 Training and testing population . . . 57
D.2 Target population . . . 59

List of Tables

2.1 OLTP vs OLAP 6
2.2 Relational vs Multidimensional databases 6
2.3 Example of confusion matrix 7

4.1 Different dataset sizes for the learning curves plotting. 27

5.1 Database tables 32
5.2 Bias and variance for each model in the Control vs MCI scenario. 38
5.3 Bias and variance for each model in the Control vs AD scenario 38
5.4 Bias and variance for each model in the MCI vs AD scenario 38
5.5 P-values of each model for each one of the three scenarios 39
5.6 Levene's test for each scenario 39
5.7 Predictive accuracy for Portuguese population 40

A.1 DICOM header tags to be treated. 49

List of Figures

2.1 Bias-variance trade off [1, 2] 8
2.2 Model accuracy on test examples as a function of the size of the training examples [3] . . 9
2.3 Hyperplane separating the two classes [4] 9
2.4 Simple example of a decision tree extracted from the work of Mehrab Sayadi, et.al. [5] . . 10
2.5 Example of instances from two classes projected onto W. [4] 11
2.6 Example of Logistic Regression [6] 11

3.1 Schematic representation of internal (I.V), external (E.V) and internal-external (I-E.V) validation. 15
3.2 Staging Alzheimer's disease with dynamic biomarkers (Image from Jack et al. [7]) 15

4.1 Full schematic view of the project 20
4.2 Example of a DICOM file using RadiAnt DICOM Viewer 21
4.3 Parallel coordinates displaying a general view of the database 23
4.4 One sunburst and two pie charts displaying relevant information about the images and patients 23
4.5 Number of patients / Age 24
4.6 Example of the table regarding some information about a patient 24
4.7 Overview of the training of the models 25
4.8 Overview of the testing stage of the target predictive models 26
4.9 Distribution by class and gender of the original population 27
4.10 Distribution of the target population by gender and diagnose 29

5.1 MRI image before anonymization. 31
5.2 MRI image after anonymization. 32
5.3 General view of the relational database model 33
5.4 Learning curves in the Control vs MCI scenario for SVM-Linear (a), Decision Trees (b), Random Forests (c), Extra Trees (d), Linear discriminant analysis (e), Logistic Regression (f) and Logistic Regression with Stochastic Gradient Descent (g) when training the models. 35
5.5 Learning curves in the Control vs AD scenario for SVM-Linear (a), Decision Trees (b), Random Forests (c), Extra Trees (d), Linear discriminant analysis (e), Logistic Regression (f) and Logistic Regression with Stochastic Gradient Descent (g) when training the models. 36
5.6 Learning curves in the MCI vs AD scenario for SVM-Linear (a), Decision Trees (b), Random Forests (c), Extra Trees (d), Linear discriminant analysis (e), Logistic Regression (f) and Logistic Regression with Stochastic Gradient Descent (g) when training the models. . 37
5.7 Confusion matrix of the target population 5.7a and the original population 5.7b 40

5.8 Precision Recall curve obtained in the target population 5.8a and the original population 5.8b . . . 41
5.9 ROC curves obtained in the target population obtained in the target population 5.9a and the original population 5.9b . . . 41
5.10 Distribution of calculated probabilities for patients diagnosed by the hospitals as Control 5.10a, MCI 5.10b or AD 5.10c . . . 42

B.1 GUI's login page . . . 53
B.2 GUI's home screen . . . 53
B.3 GUI's add new patient page . . . 54
B.4 GUI's update profile page . . . 54

D.1 Distribution of class Control by age group . . . 57
D.2 Distribution of class MCI by age group . . . 58
D.3 Distribution of class AD by age group . . . 58
D.4 Distribution of class Control by age group . . . 59
D.5 Distribution of class MCI by age group . . . 59
D.6 Distribution of class AD by age group . . . 60

Chapter 1

Introduction

Dementia is a class of diseases associated with losses of memory and thinking abilities considerable enough to interfere with the daily life of a person. Dementia associated diseases include Alzheimer's disease, Vascular dementia, Lewy body dementia, Parkinson's disease and others. According to The World Alzheimer Report 2019 [8] there are over 50 million people that have a dementia related disease or, to simplify, every 3 seconds one person is diagnosed with dementia and such frightening numbers are expected to increase up to 152 million people living with it by 2050.

The work here presented, focus on a specific dementia disease, Alzheimer's disease, representing two thirds of the total cases of dementia [9]. Currently, to diagnose such disease with total certainty is only possible with a detailed post-mortem microscopic examination of the brain [10]. The fact that, for the time being, it is not always easy to diagnose a patient with Alzheimer's disease while still alive or even at an early stage of progression does not mean that we should not discard the presence of more robust diagnostic methods to be discovered. In fact, it is possible to diagnose patients with Alzheimer's with around 95 percent accuracy by using different types of tools for the purpose. The tools that might be used to diagnose a patient are based on studying the history of the patients and their families and with that, it is then possible to assess cognitive function by neuropsychological tests. The biggest problem with such solution is that is highly dependable on medical professionals to determine the diagnose and such diagnose might take several weeks to be accomplished. In addition, the diagnose may only be performed already in a later stage of the disease when it is harder to delay or reverse the development of the disease.

1.1 Problem Description

More and more approaches based on machine learning have been used in order to develop models capable of providing an early and accurate diagnosis of Alzheimer's disease or even of a preliminary state of cognitive impairment preceding Alzheimer's at a later time of life. The biggest setback of such models is the need to guarantee their interpretability in face of the complex data available (combining imagiology, cognitive scoring exams, demography, and clinical records) and the need to guarantee their adequate generalization ability on external data i.e data the model has never seen.

At the moment, Magnetic Resonance Imaging based personalized diagnostic tools for dementia are still scarce due to the several difficulties that arise when handling such models. The acquired data to feed the models for classification is massive and heterogeneous in nature. When an MRI is performed, the output of such exam is a compilation of images displaying the brain of the patient in 3 dimensions, with a general resolution of over X thousands voxel [11]. In addiction, each image of the exam combines

medical data, with static demographic information concerning the patient and the physician involved in the exam.

Such data must be properly processed for research ends. As previously mentioned, these Magnetic Resonance images contain information about the patient and some of it must be anonymized due to the patient anonymity that must be maintained. The anonymization must be performed in ways that it makes the identification of the patient impossible to the researchers and easy for the hospital or clinic, once it receives an output from the models.

MRIs can be acquired using different technologies and protocols [12] so, it is only expectable that, before such images are handled by the models, they must be pre-processed. There are several protocols of acquisition, although, the two structural protocols of relevance for this paper are *Magnetization prepared rapid gradient echo (MP-RAGE)* and *Spoiled gradient recalled echo (SPGR)*, explained in the Background section.

1.2 Research contributions

The main goal is to develop a multi-diagnostic, multi-site, and clinically interpretable tool for early diagnosis of AD using MRI imaging initially collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and later, from several different hospitals or clinics. The solution proposed will also extend learning and assessment to new populations (new cohorts).

In accordance,the major contributions placed by our work are:

1. Validate the predictive power of the developed MRI diagnostic models for dementia in a Portuguese population using retrospective data.
   (a) Collect anonymized retrospective clinical and neuroimaging data from Portuguese hospitals of the consorcim study NEUROBIOAI.
   (b) Build a database capable of storing the images received by the hospitals.
   (c) Test existing models with such data.
   (d) Perform an external validation on the models with information the models have not yet seen.
2. Build an interface capable of visualizing important data from the database and manipulate it at will.

1.3 Outline

This work is organized as follows. Chapter 2 introduces important terms that must be acknowledge for a better understanding of the paper. Chapter 3 displays some literature review to get a know-how of the related work in the area. Chapter 4 introduces the proposed solution for this dissertation as well as the evaluation methodology used to elaborate the work. Lastly, chapter 5 presents the results of the proposed methodology followed by a discussion of said results.

Chapter 2

Background

This chapter introduces essential background on Alzheimer's Disease as well as presents information on the neuroimaging exams performed in patients, such as the heterogeneity of the exams themselves and the content of said exams, cohorts studies, or even aspects of data integration.

2.1 Medical Context

Alzheimer's disease (AD) [9] is a neurodegenerative disease affecting primarily the regions of the brain responsible for the memory of the individual such as the entorhinal cortex and hippocampus areas. AD is associated with the destruction of the neurons and their connections. Statistically, is the leading cause of dementia, comprising two thirds of cases of dementia related diseases. Cognitive complaints and mild cognitive impairments can be identified and the progress of such diagnostic can be tracked. Typically, there are three stages of importance when referring to a patient condition which are:

1. **Cognitively normal** is a state in which subjects present no signs of any kind of mental disorder such as depression, MCI or any type of dementia [13, 14].

2. **Mild Cognitive Impairment (MCI)** is a medical condition in which individuals experience memory concerns for reasons that go beyond normal aging, yet, it is not enough reason to be diagnosed with dementia as daily tasks may still be accomplished with no signs of impairment [13, 14].

3. **Alzheimer's Disease**.

Generally, the AD diagnoses are proceeded by specific exams, including Magnetic Resonance Imaging that might be either **Structural Magnetic Resonance Imaging (MRI)** or **functional Magnetic Resonance Imaging (fMRI)**. The first one, MRI, is a technique that utilizes magnets and radio waves with the purpose of photographing the inside of a body part. Such technique may aid medical staff predict if a patient will develop a disease, for instance AD or correctly diagnose it, as previously mentioned [15, 16]. The second type of MR exams is the fMRI, which is a specialized form of MRI that is used to examine the brain and, therefore, the functionality of it by measuring small changes in the flow of the blood that occur during the time frame of a specific brain activity [17]. Just like the structural MRI, an fMRI can diagnose a patient with a certain disease as well as it is commonly used to assess the impact a previous condition like a stroke had on the patient.

The aforementioned exams may differ from source to source due to technological factors and examination protocols despite the type of exam being the same, thus, contributing to the heterogeneity of exams. There are a lot of factors such as **Heterogeneous data sources and data** itself that are a

consequence of the difference between exams, for instance, the brand of the machine performing the MRI might generate an output file different from another brand. In addition, different hospitals or medical clinics might keep track of different markers when performing the exam. Furthermore, there are several protocols of acquisition of MRIs and the ones of relevance for this thesis are the following:

1. **Magnetization prepared rapid gradient echo (MP-RAGE)** is a sequence for structural brain imaging that allows for the better distinction between gray matter, white matter and cerebrospinal fluid in the brain by weighting a T1 gradient [18, 19].

2. **Spoiled gradient recalled echo (SPGR)** is a sequence characterized by superior soft tissue contrast compared with T1-weighted spin echo (SE) technique [20].

Due to these highly heterogeneous aspects in the exams, new protocols have been proposed. **Digital Imaging and Communications in Medicine (DICOM)** [21] is a protocol developed to standardize the digital format utilized in the storage and communications of images. With such protocol, Dicom files specify standard metadata to facilitate cross-source studies since the type of information in said images is the same and the format of the image itself is identical despite factors like the manufacture of the machine performing the exams.

2.2 Data Extraction

Medical image acquisition is an essential step for conducting cohort studies that support the study of diseases, along with their diagnostics, prognostics, and therapeutic treatments. In the context of neurodegenerative disorders, large-scale initiatives are available. Such images acquired are initially from **ADNI**, short for **Alzheimer's Disease Neuroimaging Initiative**, which is a large-scale cohort study conducted by a consortium of universities and medical centers to develop imaging techniques and biomarkers procedures in pursuance of early detection of Alzheimer's Disease and to keep track of the development of the same [22]. Later on, images are also acquired from partner hospitals or medical clinics where they are stored in the respective **Picture Archiving and Communication System or PACS** systems of the entities that supplied the images. **PACS** is, as the name implies, software to store and facilitate the communication between medical centers for imaging [23]. In the latter process of acquisition, **Web scraping** is necessary to perform the task. Web scraping is a technique used to extract certain desired data from a specific website, thus simulating a task that would be performed manually by a human. The automatic extraction of unstructured data into structured databases, designed as the developer needs, allows for the use and manipulation of the structured data as the user sees fit [24]. However, before such technique is performed with the purpose of extracting images, the anonymization of the Dicom files is of the utmost importance since **Protected Health information** or health data from an identifiable patient is kept secret and compliant with the laws protecting the privacy of a subject [25]

After the extraction of the images, subsequent mining steps may be conducted to learn descriptive and predictive models. This learning step can be generally led by the discovery of **Neurodegenerative biomarkers** that are biological indicators of a medical state that can be objectively measured and be of help when studying, tracking, and predicting outcomes for a specific neurodegenerative disease. [26]. Such models aim to aid in the process of diagnosis and aid in assigning an early intervention for a possible disease a patient may come to develop. To perform such a demanding task those models must yield good guarantees of predictive accuracy.

2.3 Data Storage

During the process of decision and elaboration of a database, it is required to assess the main functionality/purpose of the system to be developed. These systems might be differentiated into two main categories, transactional or analytical systems.

2.3.1 OLTP and OLAP Systems

OLTP stands for On-line Transaction Processing, therefore being the pilar of transactional systems, as mentioned before. This type of system is known for the simple on-line transactions (Update, Insert, Delete) that performs at a high volume of requests.

On one hand, the main focus of OLTP is to process the queries at hand as fast as possible while at the same time maintain the data integrity in the database, this is, the multiple accesses from several users does not jeopardize the correct value of the data. On the other hand, OLTP is susceptible to security problems. Since data is stored in full, it becomes susceptible to theft from external entities. Furthermore, a single failure, an input mistake, for example, might cause a chain reaction where it becomes a highly demanding task to recover from the original failure, therefore, resulting in high costs of time and money.

OLAP stands for On-line Analytical Processing and as the name suggests it is the foundation of analytical systems. This type of system is characterized by a low volume of transactions. The consultation of data in such systems might turn out to be a complex task and the data must be aggregated in order to be displayed due to the high level of summarized data in the database so, it is only normal to expect that for a specific use case, several data must be aggregated for an analytical report.

Contrary to OLTP, where the measure of performance is calculated by the total number of transactions per a specific timeframe, OLAPs measure of performance is the efficiency of the query at hand since the output of such request is then used for analytical purposes such as Data Mining where correct data is a requirement. Wrong output from the queries would lead to wrong assumptions when analyzing a certain situation. This way, it is possible to conclude that the time a query takes to be executed is not a priority, rather than, the correct calculation of its output.

The OLAP systems are also known for the inclusion of a dimension called Time. With such dimension in play, the stored data allows for the reporting of a subject´s history throughout time.

2.3.2 Relational and Multidimensional Databases

A relational database, as the name suggests, is based on relational models. These types of databases organize data in rows and columns, in a 2-dimensional form,

Such kind of solution has the ability to ensure that the data is consistent all through every instance or application. Furthermore, it is also labeled ACID which stands for Atomicity, Consistency, Isolation, and Durability since it handles the data at a granular level.

On the other hand, we have this type of database that is considered the next step to relational databases. Is common to see multidimensional databases built with relational ones. As the name suggests, this type of database contains multidimensional arrays with 3 or more dimensions whereas, relational databases contained arrays with 2 dimensions.

Multidimensional DBs are optimized for OLAP applications and/or data warehousing. Multidimensionality allows the developer to handle data as he/she sees fit. We might even add a time dimension to keep track of the history of a subject/object.

To best assess which solution might be better for the task at hand, table **??** displays some pros and cons about each type of solution regarding some quality attributes.

	OLTP	OLAP
Data origin	OLTP is the original source of data	Data in OLAP is a result of the consolidation of several OLTP databases
Objective	Consult/view the data. Performs tasks as fast as possible	Analyse the data. Elaborate statistical analysis in order to aid in the decision making process. Integrate different data sources for the creation of a consolidated database.
Performance	In general, it is very fast due to the high detail of data stored	Optimized for analysis and data reading. Queries might take longer to run depending on the amount of data
Modifiability	Data is volatile, therefore, modifiable	Data is referring to a certain time frame, so data is not volatile
Maintenance	Data updates are performed during each transaction. High number of updates. Data backup is crucial since errors might develop chain reaction errors. Loss of data incurs in high costs.	Data is updated during its loading. Might be performed periodically. Backup regularly is not necessary. Reload of the OLTP databases to recover data.
Queries	Simpler and standardized. Inefficient for great amounts of data	More complex, requires aggregation of data
Data Structures	High level of detail	High level of summarized data. Structured in several dimensions
Security	Users can freely manipulate data	Users are only allowed read and insert data

Table 2.1: OLTP vs OLAP

	Relational Database	Multidimensional Database
Data origin	OLTP is the original source of data	Data in OLAP is a result of the consolidation of several OLTP databases
Objective	Consult/view the data. Performs tasks as fast as possible	Analyse the data. Elaborate statistical analysis in order to aid in the decision making process. Integrate different data sources for the creation of a consolidated database.
Performance	In general, it is very fast due to the high detail of data stored	Optimized for analysis and data reading. Queries might take longer to run depending on the amount of data
Modifiability	Data is volatile, therefore, modifiable	Data is referring to a certain time frame, so data is not volatile
Maintenance	Data updates are performed during each transaction. High number of updates. Data backup is crucial since errors might develop chain reaction errors. Loss of data incurs in high costs.	Data is updated during its loading. Might be performed periodically. Backup regularly is not necessary. Reload of the OLTP databases to recover data.
Queries	Simpler and standardized. Inefficient for great amounts of data	More complex, requires aggregation of data
Data Structures	High level of detail	High level of summarized data. Structured in several dimensions
Security	Users can freely manipulate data	Users are only allowed read and insert data

Table 2.2: Relational vs Multidimensional databases

2.4 Data Science Concepts

This section is intended to provide some background on important Data Science concepts that the reader must understand in order to better acknowledge the work displayed in this dissertation.

2.4.1 Predictive assessment

In data science, predictive models typically correspond to classification or regression models depending on whether the condition of interest is nominal or numerical. Since an objective of this work is to predict a

discrete class output, we map the target task as a classification model. **Classification** is the prediction of the most suitable class for a data observation given as input. For a set of data observations X **=** $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ and a set of attributes Y=$\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m\}$, a model M maps observations from X to discrete output variables Y, thus predicting the adequate class for $\mathbf{x}_{new}$ [4]. The classification process of the developed models is intended to differentiate and clearly diagnose a patient as CN or with MCI or AD.

In order to assess how models behave in terms of generalization ability, several measures are used to evaluate the performance of a model. Since the model at hand is a multi-class classifier i.e. cardinality of the output space is $>$ 2, it is required to use metrics capable of measuring the performance of the different classes.

The first technique to be applied is a confusion matrix since it is a table summarizing the performance of the model by displaying the contrast between the predicted values and the real values as seen in Figure 2.3. To further analyze it, we have to find the TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives) for each class. To simplify, consider the example of HC (Healthy control) in which, TP = a; TN = (c+f+h+i); FP = (b+c); FN = (d+g).

		Real diagnosis		
		HC	MCI	AD
Predicted diagnosis	HC	a	b	c
	MCI	d	e	f
	AD	g	h	i

Table 2.3: Example of confusion matrix

With the confusion matrix calculated, it is possible to calculate the performance metrics *Accuracy, Precision, Recall and Specificity.* Simply calculating the *Accuracy*, the ratio between the number of correctly classified points to the total number of points, is not a good approach since the data could be highly imbalanced so the model classifies all the data points as the majority class data points. Since *Accuracy* might not be a good metric when facing imbalanced datasets, it is introduced the following metrics for the example where given a specific class of interest, for instance, HC, then:

1. **Precision** is the fraction of instances correctly predicted as HC out of the total classified instances as HC,
$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.1}$$

2. **Recall or Sensitivity** is the fraction of instances correctly predicted as HC out of the the total HC subjects,
$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.2}$$

3. **Specificity** is the fraction of all negative instances that are correctly predicted as negative,
$$Specificity = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.3}$$

4. **Balanced Accuracy** is a better metric to use with imbalanced data since it accounts for both the positive and negative predicted classes without misleading with imbalanced data,
$$BAC = \frac{\text{Recall} + \text{Specificity}}{2} \tag{2.4}$$

By calculating such metrics under different probability thresholds, it is then possible to calculate **Area under the ROC (receiver operating characteristic) curve**. ROC is the graph showing the performance

of a classification model at all classification thresholds in where the X-axis is the Specificity and the Y-axis is the sensitivity. The area under the ROC (AUC) represents how well the model is capable of distinguishing between HC, MCI, and AD. The higher the AUC is, the better the classifier is at predicting each class of interest.

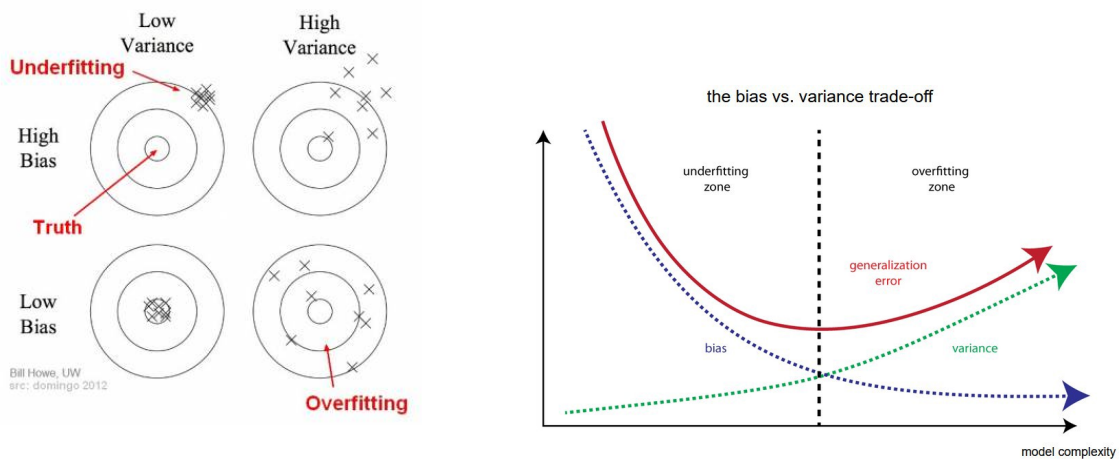
As it is possible to see from the aforementioned information, Recall and Specificity are asymmetric as they only consider either the positives or the negative values but on some occasions, it is required to consider both values. For that reason, a metric to be introduced is the **Matthews Correlation Coefficient** or MCC which is a symmetric metric. The higher the coefficient is the better the predictions are despite the fact that one class might be disproportionately under or over-represented.

$$MCC = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{2.5}$$

Another task to perform when performing the external validation is to detect if the models are overfitting or not. **Overfitting** occurs when a function corresponds too closely to the training data, i.e. the model does not generalize well from the training data, performing poorly on external/new data.

To detect such an event, evaluation metrics from the training set and the test set will be compared to assess how close are the two values for a varying number of observations. In the event of the model presenting a high accuracy in the training set and a considerably lower accuracy on the test set tell us that the model is overfitting, if the train accuracy remains constant and the test accuracy worsens, and external data will not be properly classified.

Another way to detect if the model is more susceptible towards underfitting or overfitting risks is by looking at the **bias-variance trade off**. **Bias** represents the difference between the predicted values and the real values. **Variance** is the variability of the predictive model for a given data point. Figure 2.1 illustrates the bias-variance trade-off. A model with a high bias does not pay enough attention to the training data and simplifies the model to a point where it is oversimple (underfitting). On the other hand, a model with high variance excessively analysis the training data and does not perform well on new data due to the low generability.



(a) (b)

Figure 2.1: Bias-variance trade off [1, 2]

The last evaluation principle worth mentioning is a **learning curve**, which provides an overview of predictive performance considering different sizes of training samples. According to Perlich et.al.[3], learning curves are commonly used to display the predictive accuracy of the models on the test examples

considering the variation of the training examples as shown in Figure 2.2

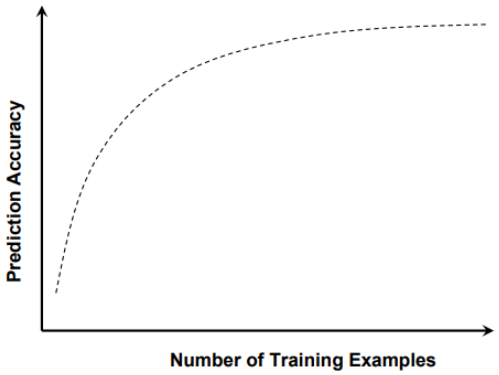


Figure 2.2: Model accuracy on test examples as a function of the size of the training examples [3]

Perlish et.al. [3] also acknowledge the different use of learning curves in Machine Learning. These may be used in two different scenarios:

1. In most cases, learning curves are used to obtain an overview of the predictive generalization performance regarding different training sizes as is the case in this thesis.
2. More specifically in Artificial Neural Networks, learning curves have been used to show the differences between in and out-of-sample performance considering different training sizes.

2.4.2 Predictive modeling

Part of the work presented in this thesis is based on predictive models developed prior to my involvement in the external validation. For that reason, this section is meant to aid the reader by providing several insights on machine learning models that appear further ahead.

The first model being presented is the ***Support-Vector Machine*** or *SVM* for short. According to Cristianini [27], *SVM* is a supervised machine learning model that can be used for either classification or regression problems and makes use of algebraic and statistical properties to distinguish two classes. *SVM* classifies a new instance by drawing a hyperplane between two classes.

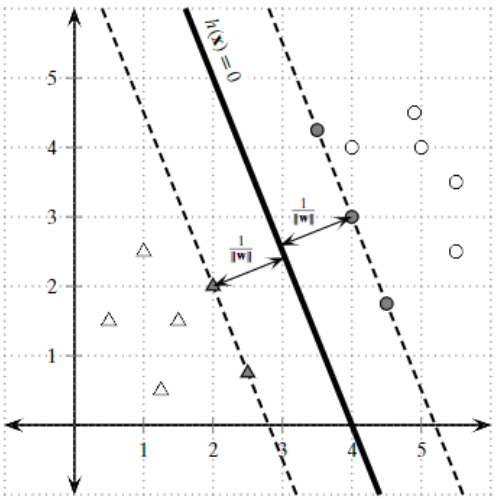


Figure 2.3: Hyperplane separating the two classes [4]

There are cases where a linear division is not possible. In such cases which it is not possible to simply divide the two class instances, *SVM* makes use of the *Kernel trick* which projects the alinear data into a dimension with more coordinates until it finds one where it is possible to draw a plane capable of correctly divide the class instances [28].

Similar to *SVM*, ***Decision Trees***, *DT* for short, are supervised learning algorithms that may be used for either classification or regression problems. As the name implies, a *DT* is in fact a tree that displays the features that best discriminates the observed population, and the deeper we go into a *DT* the higher the level of detail.

In order to Figure out the root of the tree, there are several algorithms capable of doing so. An example is the classic ID3 algorithm, which uses information theoretical measures to assess the relevance of attributes. The lowest entropy is selected to be the root and so on until a decision tree is created. An example of a *decision tree* can be seen in Figure 2.5 where the goal is to predict if a person is either healthy or diabetic.

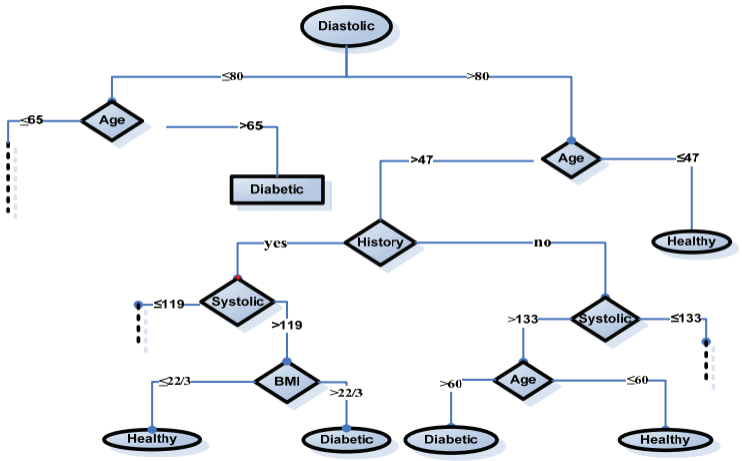


Figure 2.4: Simple example of a decision tree extracted from the work of Mehrab Sayadi, et.al. [5]

When mentioning ***Random Forests or Extra Trees***, *RF* and *ET* respectively, one is mentioning an ensemble of *decision trees*. The idea is that the more *decision trees* the ensemble has, the more accurate the model will be as the output of the models represents the output of several *decision trees* combined [29, 30]. The differences between each of the two predictive models are:

1. Both models split the training data but such split is performed in a different way since *RF* looks for an optimal split point in the training data whereas in the case of *ET*, said split point is chosen at random.

2. While *RF* uses the ”Bootstrap Method” [29] to calculate the quantity of data sample to use, *ET* uses the whole entire data sample.

Another predictive model worth mentioning is the ***Linear Discriminant Analysis***, *LDA* for short, whose goal is to find a vector $\boldsymbol{w}$ that maximizes the separation between the classes $\mathbf{y}$ of labeled data consisting of n-dimensional points $\mathbf{x}_i$, after projecting said classes onto $\boldsymbol{w}$.

Similar to ***Principal Component Analysis (PCA)***, *LDA* looks for linear combinations of variables that best describe the data [31] but contrary to *PCA, LDA* deals with labeled data and tries to maximize the discrimination between the classes.

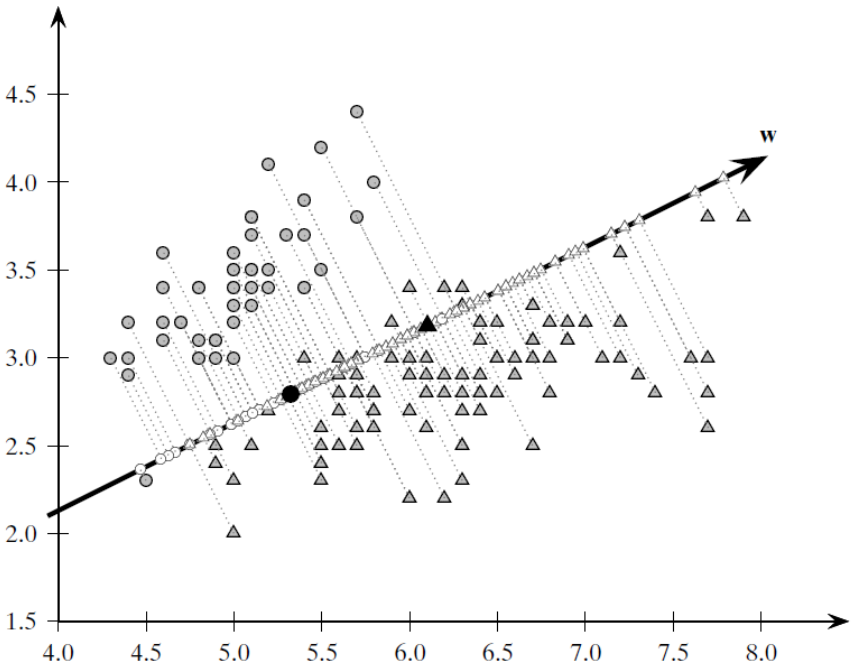


Figure 2.5: Example of instances from two classes projected onto W. [4]

According to Shwartz et al. [32] ***logistic regression*** or *LR* for short, is a classification algorithm that is used to assign observations to a discrete set of classes. A LR is a specific case of a perceptron model, where the activation function is given by the sigmoid function,

$$\phi_{sig}(x) = \frac{1}{1 + e^{(}- x)} \tag{2.6}$$

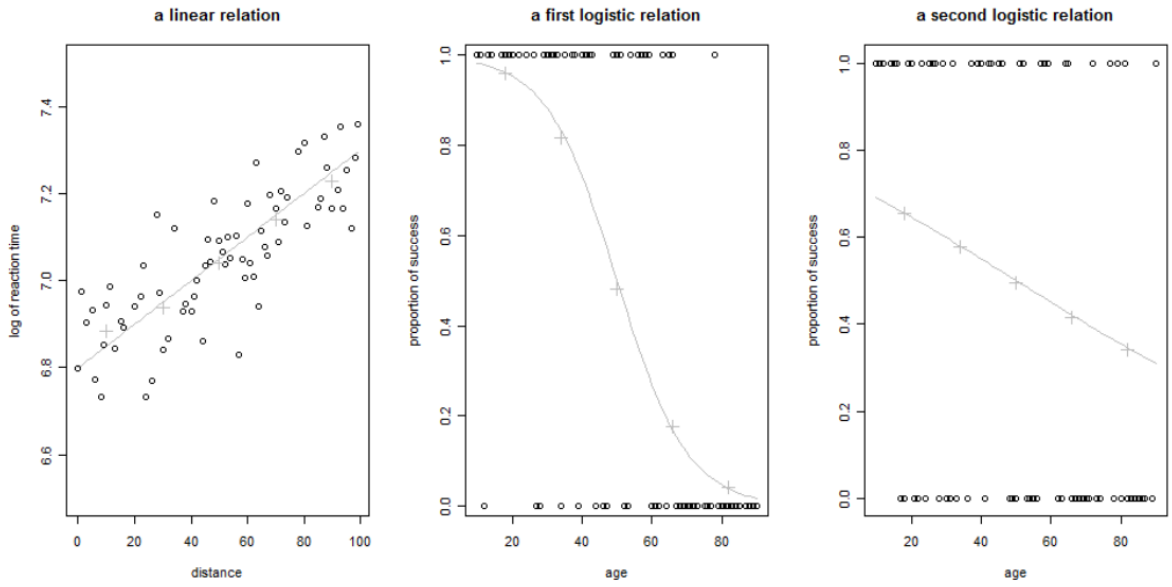


Figure 2.6: Example of Logistic Regression [6]

Chapter 3

Related Work

The consolidation of data and its external validation are not novel topics in computer science, although, the purpose of the application domain and the unique challenges associated with the available data sources make a given project unique based on its own constraints. Therefore, the purpose of this section is to present a compilation of related work. Section 3.1 introduces work related to data consolidation whereas section 3.2 presents work referring to external validation.

3.1 Data Consolidation

Volosnikov et al. [33] proposed a tool capable of allowing unified access to heterogeneous and distributed data. According to the paper, the heterogeneity of data sources increases the difficulty to perform comprehensive research. Furthermore, the data presenting the heterogeneous indicators of medical exams range in type, some even might be considered heavy, for instance, MRI or fMRI. Such images demand an intensive preprocessing phase in order to use them in a research analysis. To tackle the problems that arise due to the heterogeneity of data sources and the the required preprocessing of the images, the committee of authors then introduces the developed tool. Such implementation makes use of a service-oriented architecture, commonly known as SOA, preventing a series of problems that, otherwise, would have arisen. Compliance to the law when it comes to handling the personal information of each patient subjected to the study along with scaling difficulties and the use of new resources are examples of problems could have appeared. The tool developed uses python libraries to access and store the heterogeneous data and the interface and work environment of the tool was implemented using MEAN stack or Mongo, Angular, Express, Node.

Similar to the work previously mentioned [33], the solution presented in this paper must handle the consolidation of medical imaging and all the problems that may emerge with it. The data used to preform our own analysis is handed by a grid of hospitals hence the heterogeneity of the data sources. Since the files gathered and being used are raw MRIs, it is also necessary for them to undergo a preprocessing phase in order to analyse them. Since each hospital is a unique case, each one of them demands a different way to extract required data from the servers into our workstation. For the time being, web scrapers are being used to access and extract data from the PACS system of each hospital. The interface of the solution here proposed is implemented in python as for all the access and storage of content in the database created in our workstation. Since the anonymization of data is performed by the hospital by a script developed alongside the proposed solution, compliance to the law in terms of handling personal information does not raise a problem since the anonymization process was accepted by each supplier of data and the research work is compliant with HIPAA, GDPR and other data privacy regulations.

Data Warehousing, as the name suggests, it is used to store data from disparate sources. The work of Saliya Nugawela[34] identifies the main obstacles of data integration of healthcare data and the proposal of a data warehousing model capable of integrating fragmented data in a cardiac surgery unit. The work proposes a star schema to organise the data collected along with an enterprise architecture. The main difference between such solution and the one presented in this thesis, is that the solution presented here follows a snowflake schema. The less space it is wasted, the more information can be stored. In a star schema a lot of the information turns out to be redundant whereas in the snowflake there is almost no redundancy.

3.1.1 Data security and privacy related issues

With the use of digital technology in practically every aspect of medicine to collect clinical data related to a patient, personal information is stored inside each digital files. As previously mentioned, Dicom entered the field to simplify the exchange and storage of digital data. According to Newhauser et al. [35], researchers have the obligation to remove PHI from all electronic medical records used in the research of a patient outcome for a certain condition, specially, before such images are made public. Attributes inside a Dicom image might be deidentified, pseudoanonymized or even fully anonymized in order to comply with the regulations of HIPAA [36] in the United States of America or, for instance, GDPR [37] in Europe. The goal of the article was to develop and test a prototype software code capable of automatically anonymizing all the required information stored inside the EMRs without losing the records integrity, thus, proposing the deletion or overwriting of certain dicom fields containing protect health information.

Maintaining data security without leaving an electronic medical records like Dicom images opened to privacy related issues tends to be a rather delicate and unique process. The way each image is processed highly depends on the objective of the recipients research, therefore, depending on the purpose some information might be necessary and can not be entirely anonymized. Throughout the years, several works [38, 39, 40, 41] have been discussing different methods and techniques for the anonymization of the PHI present inside the Dicom images and the method presented in this article is no different as explained further ahead.

3.2 Predictive model validation

After the development of a predictive model, it is necessary to validate it in order to understand how the model behaves in terms of performance. A core element of validation of the predictions models is to contrast internal with external validation. Section 3.2.1 presents some related work expressing the usefulness of biomarkers for aiding in an MCI diagnosis when there is suspicion of said condition. Despite the fact that section 3.2.1 is not the main focus of the work, it still has some considerable synergies with the work at hand.

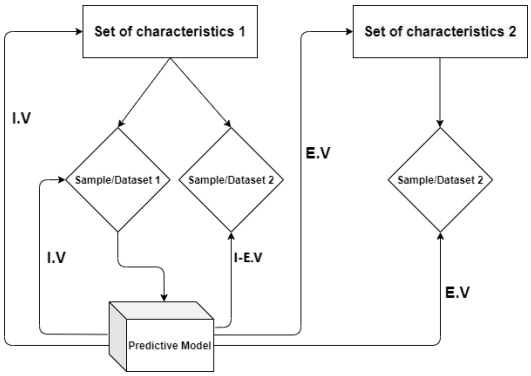


Figure 3.1: Schematic representation of internal (I.V), external (E.V) and internal-external (I-E.V) validation.

3.2.1 Biomarkers Discovery

Bocchetta et al. [42] studied the relevance of AD biomarkers such as cerebrospinal fluid (CSF), medial temporal atrophy (MTA), fluorodeoxyglucose positron emission tomography (FDG-PET) and amyloid-PET by AD European consortium centers, obtained by inspecting MRIs in the diagnosis of MCI. According to the article, the most used biomarker is clearly MTA with 75% of the respondents claiming to always or at least frequently use it. The second most used is CSF markers with 22% of respondents using it, followed by FDG-PET with 16% and finally amyloid-PET with 3%. In terms of confidence in the use of such markers in the early diagnosis of MCI, only 45% of the consortium centers that answered the survey considered that MTA had a ”moderate” contribution to the diagnosis whereas 79% felt ”very/extremely” confident in a diagnosis of early MCI due to AD when levels of amyloid and neural injury biomarkers were abnormal, especially when the measurement of the levels of both were simultaneously abnormal, thus, being an indicative of AD signature.

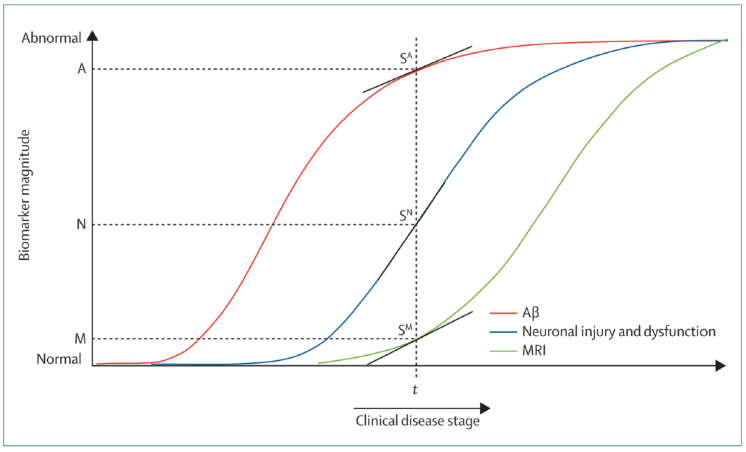


Figure 3.2: Staging Alzheimer’s disease with dynamic biomarkers (Image from Jack et al. [7])

Other literature corroborates conclusions as the ones, aforementioned, for instance, the work by Jack et al. [7] in which, the authors provide a framework developed with the purpose of testing hypothesis presenting correlations between changes in AD biomarkers throughout time and clinical diseases stages or even between temporal changes in AD biomarkers themselves. As shown in figure 3.2, it is possible to understand that biomarkers like MTA, detected through the use of structural MRI might not be as relevant as one would predict despite the frequent use of said markers due to abnormalities only presenting at a later stage of the disease. On the other hand β-amyloid abnormalities seem to appear at an early stage

of the disease, thus, corroborating the highly confidence level in a diagnosis where amyloid levels were abnormal.

3.2.2 External validation

Despite the fact that a predictive model is validated internally, an external validation is required and essential since, in that way, it is possible to test the model on a population acquired in an independent way. By doing so, external validation allows for the assessment on the generalization of a predictive model, allowing for a better understanding on how the model performs on a new population.

Most of the predictive models used in a Alzheimer's disease related issues make use of deep learning techniques. According to the work of Qiu et al. [43] there is a lack of external validation methods being implemented in deep learning techniques based predicted models since such models are developed, i.e. trained and tested, with data from a single group of subjects who share a defining characteristic. The fact that a lack of external validation methods exists, deep learning models applied to AD tend to fall short on the expected outcome considering the fact that such models have a decrease on performance and their comprehensibility is limited since these models work as a "black-box" and provide no elucidate diagnostic review.

Furthermore, external validation is necessary in prediction research. The work of Bleeker et al. [44] elucidates the fact that predictive models tend to perform better when facing data used to train and develop the model rather than when facing data new to the model. The results from predictive models tend to be considered with regard to the internal validation and with almost no regard for the external one. Bleeker et al. [44] present the limitations to internal validation, therefore, expressing the importance of external validation. The predictive model used in the paper aims at classifying the presence of serious bacterial infections in children with fever (total amount of 376). Internal evaluated performance on average of 0.83 for the apparent area under the receiver operating characteristic curve and 0.76 after applying a bootstrapping method to provide bias-corrected estimates of model performance. After validating the model internally, a small set of 179 individuals was validated externally and the authors obtained a performance of 0.57 proving that only validating a small data set internally is not enough and in the future models who do it, tend to fall short on performance. External validating is, therefore, considered essential and vital to be performed on a model before inserting it in clinical practices.

To summarise, let us consider the work of Siontis et al. [45] where the goal of the authors was to evaluate how often newly developed risk prediction models undergo external validation and how well they perform in such validations. The method used to try and find an answer was to evaluate 127 new prediction models. Only in about 25% of the models, an external validation was encountered and that the probability of having such validation method to be performed by different authors was 16% proving that external validation of predictive models in different studies is uncommon and, therefore, their performance might be considerably lower when facing said validation.

To perform a clear external validation on a predictive model, it is necessary to expose such model to different data, that it has not encountered before. The difference in the data has to be, according to Moons et al. [46], in these parameters:

1. Temporal differences so that a temporal external validation might be performed since the individuals presented on the data that the model is facing belong to the same cohort but to different time periods.

2. Geographical difference in the data allow for a geographical external validation considering new individuals from different locations, this is, patients subject to prediction by the model are from a different clinic or hospital.

3. Domain differences express new individuals who are considerably different from the individuals from which the model was developed representing, thus, a domain validation.

The procedure, then, consists in applying the models to the data with the aforementioned differences and recalculating the performance of the model based on discrimination, calibration and classification measures.

3.2.3 Internal and internal-external validation

Depending on the literature and on the predictive model case where such model is inserted, external validation may or may not be essential to correct the model due to low values in perfomance. However, internal validation and, in some cases, internal-external validation are some types of validations that are present in the developmento of the models. The work by Steyerberg et al. [47] expresses the fact that internal validation is essential and the preferred method for validation is the bootstrapping aproach to estimate the performance of the model. Since some type of external validation might be considered in time of development, the authors also recommend an internal-external validation. That way, the model is tested with a different sample, although with the same characteristics, as seen in figure 3.1, keeping the model from returning overly optimistic performance values, thus offering a more realistic assessment.

Chapter 4

Development

To better understand the solution at hand, first, we need to acknowledge the functional requirements of the said solution. Such requirements aim to represent what the developed solution must be able to satisfy and how it performs a certain task given a specific input by the user. Depending on the goal, different tasks must be performed to obtain the correct output for the user query. For that reason, the objectives of this dissertation are:

1. Guarantee proper anonymization and privacy of neuroimaging data.
2. Receive imaging from the hospitals/clinics.
3. Easily and quickly handle input from users.
4. Store required data in a database.
5. Return analysis based on the individuals from an hospital e.g. diagnosis.
6. Perform external validation on the new population showing the generalization guarantees and vulnerabilities.

For the better understanding of the reader, it is important to note that the solution here presented was developed with the purpose of aiding the Institute of Biophysics and Biomedical Engineering under the project NEUROBIOAI. The need for a solution capable of storing the data acquired from the partner hospitals as well as a critical analysis of how the predictive models behave under a new population, the Portuguese one, resulted in the solution presented in the next sections.

4.1 Project's Architecture

A scheme of the architecture of the tool developed as well as other relevant steps can be found in Figure 4.1.

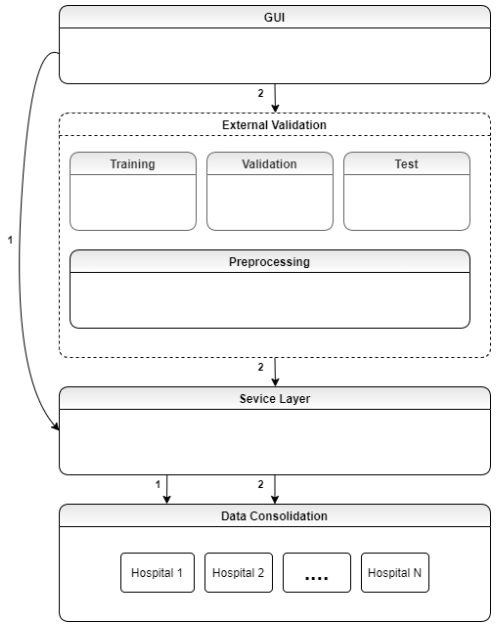


Figure 4.1: Full schematic view of the project

As seen in the scheme, the project consists of four main subgroups:

1. **Data consolidation** - The section consists of acquiring and handling images from the partner hospitals related to the patients at a specific hospital. Such images must be anonymized and, later, stored in order for the predictive models to have access to this new information. Such data is consolidated in the developed database.
2. **Service layer** - This layer is meant for handling all sorts of requests to access and alter the database if necessary.
3. **External validation** - The models, after proper training, must undergo a critical analysis so that it is possible to assess how the models handle new data.
4. **Graphical user interface** - In order for everyone to have access to the database, a centralized app (GUI) was developed in order to manipulate the database as the user sees fit.

Figure 4.1 shows two possible pathways: 1 and 2. Path number one is the one where the user must request the classification of a patient outside the developed GUI. And path 2 is the one where the GUI is capable of requesting that classification.

4.2 Data Consolidation

For a better classification of future patients, the predictive models need not only images from ADNI but also require images from hospitals or clinics. The preprocessing and spatial alignment on new data is essential to make images more easily comparable, but not necessarily similar. In that matter, before receiving such images, it is necessary to prepare them and, only after, extract such data from the partner hospitals and clinics.

4.2.1 Data anonymization

The first stage of the extraction of the images is to anonymize the information that might be identifiable of the patient. Inside each DICOM file, besides the image itself, there are several tags with information

regarding the patient as seen in 4.2. The main goal is to de-identify or remove data from the DICOM files from the hospitals/clinics, thus, enabling the sharing of such images to outside of the hospital guard without breaking any security and data privacy protocols.

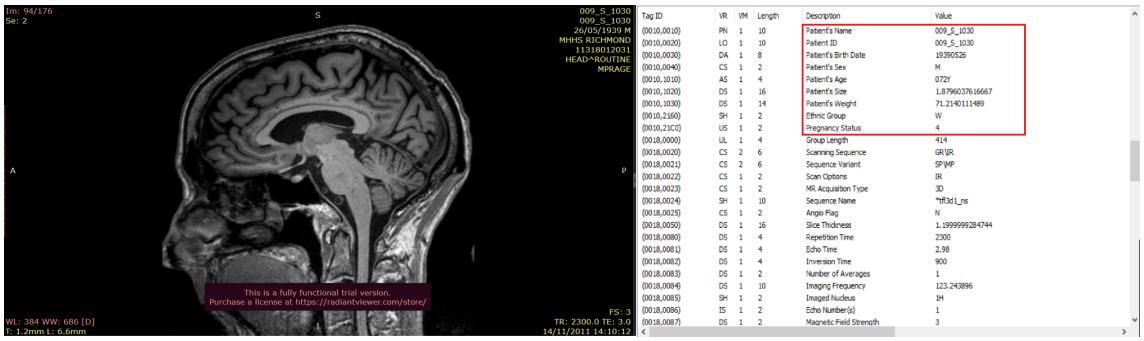


Figure 4.2: Example of a DICOM file using RadiAnt DICOM Viewer

With this process, the data is anonymized to the entity receiving the images, as it is not possible for the receiving end to obtain the original values from the images or find the original person behind the anonymization as the data is either removed or de-associated from the patient as a consequence of using hash keys to replace ids. The data is de-identified to the hospital/clinic end since the alterations to the image are stored, thus, enabling the hospital/clinic to identify the patient once a diagnose has been made by the receivers end (IBEB).

The process of anonymization consists in deleting or replacing with random values DICOM's header tags that allow for identification of the patient. Tags that are anonymized are permanently de-identified from their source. On the other hand, de-identification of some tags replaces the tags' values with artificial identifiers, random key of 10 characters, that can still be used to re-identify the patient, but only by authorized personnel of the hospital sharing the data.

Authors in different literature [48, 49, 50] provide different advice regarding the removal of some of the tags kept within the scope of the research project. It must be emphasized, though, that the tags from table A.1 are either de-identified – and only re-identifiable by the hospital - or completely anonymized as their nature does not allow them to be used in patient re-identification efforts. The decisions regarding the anonymization of the images are considered and thought of under the hospitals/clinics supervision. Before implementing such a script, the project's partners must accept and agree on the process explained above.

4.2.2 Data Storage

After the extraction of the images from the hospitals, it is required to have someplace to store the content of said images. For that reason, a local database must be created with the required tables to store the data that feeds the models for classification.

With all displayed in section 2.3, the solution that fits best the requirements is a relational model due to several reasons such as:

1. All the information can be stored in a single database so, OLAP functionality would not be that much of an asset.
2. Since one of the main reasons is for the scientists at IBEB to consult the data as it is stored in the database without any integration performed to it, a relational model suits the problem better.
3. Each image contains a high amount of data and that amount must be multiplied by hundreds of thousand other images, OLAP queries would take substantial time to run.

4. Future users of the database are not experienced in this matter so there is a need for a simple, efficient, and free way of inserting and manipulating the data.

5. Several data that was not relevant to the project was discarded in the data anonymization process so, the data being stored is of the utmost importance and must not be summarized as the user may need to see the raw content of each entry.

Although a multidimensional approach would benefit the project, after careful consideration and since the database would run at a local level with limited access, a relational approach was a more suitable way to store the data as it also leaves space for a multidimensional approach in the future if the project has such necessities, through the use of a ROLAP (Relational On-line Analytical Processing) method creating a new layer on top of the relational one.

4.3 Graphical user interface

The primary goal taken into account during the development of the GUI was to allow the user to insert new data into the database without having to write any SQL query in the console. The GUI allows for the easy and fast insertion of new patients into the database with very few interactions or effort.

To sum up, the GUI must perform the following requisites:

1. Add new patients by providing the age of the first visit of the patient at the hospital and the preliminary diagnosis. When adding a patient, it is also possible to select the Dicom images of the said patient from a directory.

2. Access and display the tables with the data from a patient or all patients, among other relevant data.

The GUI was developed in python with several libraries, including Dash, the framework where the interface is built on. Some images of the Dash app can be seen in the Appendix B.

4.3.1 Input files

As previously mentioned, the main goal of the GUI is to allow for the insertion of new patients. The hospitals and clinics send a CSV file with minimal information regarding the patients together with all the DICOM files concerning the patients' MRI.

For that matter, the developed app is prepared for receiving simultaneously a CSV and all the images the user wants to. The service layer, then, checks the database for duplicates and in case it finds, it does not insert the content of said MRIs into the database.

4.3.2 View Data

Not only the insertion of new data was considered in the development of the database. One great asset of the Dash library is that, since it uses the Plotly library, it is capable of displaying several visualization tools that able the user in terms of getting to know the population present in the database.

In order for a better understanding, the figures ahead display the database populated with only a few patients.

Figure 4.3 displays a parallel coordinates chart capable of displaying vital information about the subjects such as age, gender, the hospital where the images were taken, and the diagnosis attributed to

the patient. This visualization tool also represents the lines in such a way that it displays the diagnosis by color.

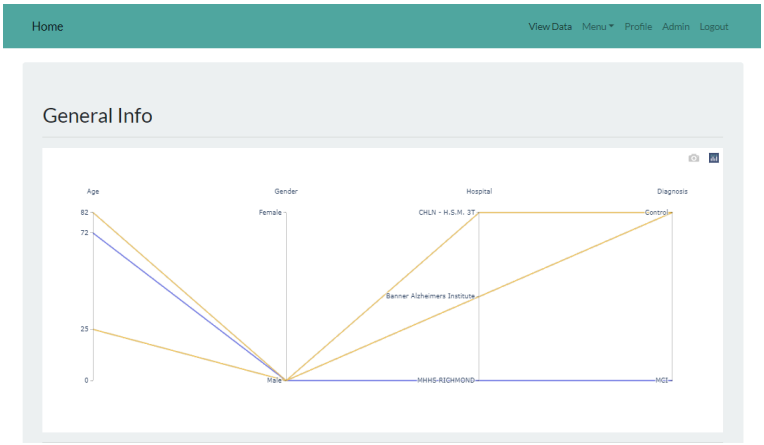


Figure 4.3: Parallel coordinates displaying a general view of the database

One great concern was to enable the user to quickly get statistical values regarding different aspects of the database population. Figure 4.4 shows data about the Diagnosis gender, imaging protocols of acquisition, and the image source.

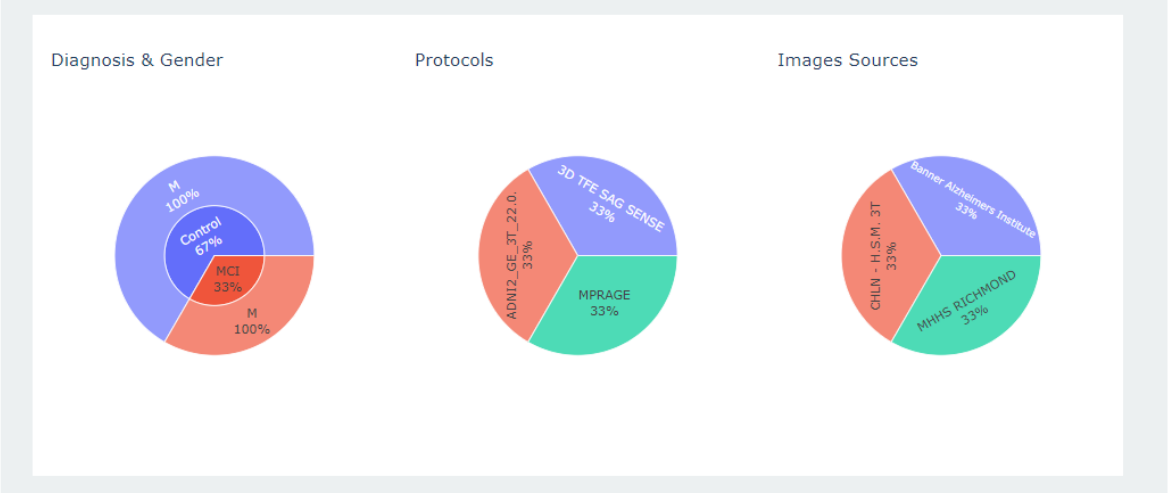


Figure 4.4: One sunburst and two pie charts displaying relevant information about the images and patients

The user is also able to get an idea of how each class of interest is affecting the subject of several ages. The scatter plot in Figure 4.5 shows, for each age, how many patients there are with the a certain diagnose.

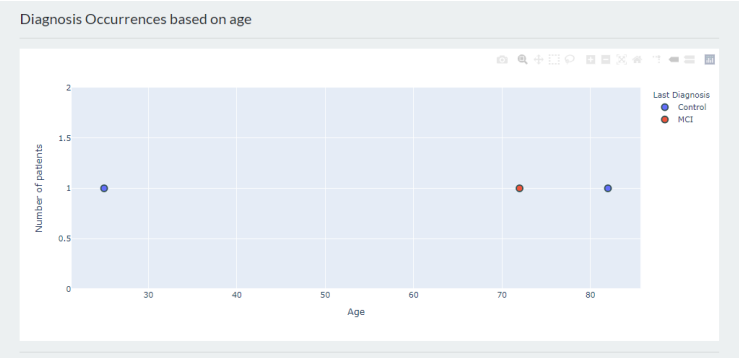


Figure 4.5: Number of patients / Age

Lastly, the user is able to consult all the tables as he/she sees fit as the example on Figure 4.6 shows. The only table that is not being displayed is the table regarding the usernames and other details of each user profile in the database. Only a user assigned with such privileges can access such data.

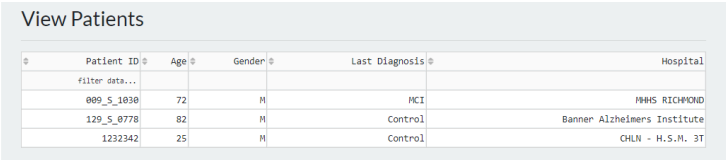


Figure 4.6: Example of the table regarding some information about a patient

4.4 Models Description

As previously mentioned, part of the work presented in this dissertation is based on predictive models developed prior to my involvement in the project. Such predictive models were developed by the investigator Vasco Sá at IBEB and so were the illustrative images used in section 4.4. In order to better understand the work presented next, let us first start this section project by clarifying the reader on how the existing predictive models behave and how they are developed.

4.4.1 Training the models

When mentioning *"the models"* in any part of this thesis, the reader must understand that said models are the result of the output of seven trained models which are: Support-Vector Machine (SVM), Decision Trees (DT), Random Forest (RF), Extra Trees (ET), Linear Discriminant Analysis (LDA), Logistic Regression (LR), and Stochastic Gradient Descent (LR-SGD).

In the training process of each of the seven models, it is performed a five-fold Cross-Validation. With this, feature selection is also performed in order to reduce dimensionality. The classification of an MRI is based on the previously mentioned models along with a Genetic Algorithm Hyperparameter Optimization.

After each model has performed the aforementioned steps, a ranking metric is calculated in order to assess the best models to use in the test set. Such ranking metric is as follows:

$$\textit{Ranking Metric} = \overline{\mathbf{MCC}}_{testfolds} - \sigma\mathbf{MCC}_{testfolds} \tag{4.1}$$

where $\overline{\mathbf{MCC}}_{testfolds}$ is the average value from the Matthew Correlation Coefficient of each fold and $\sigma\mathbf{MCC}_{testfolds}$ is the standard deviation of the Matthew Correlation Coefficient of each fold.

After calculating a ranking metric for each model, it is then necessary to choose the ones to use in the test set based on said metric by selecting the ones that present a ranking metric higher than the average of ranking metrics of the seven models.

After all these steps it is time to refit the models on the entire training set. Figure 4.7 presents an overview on the aforementioned process.

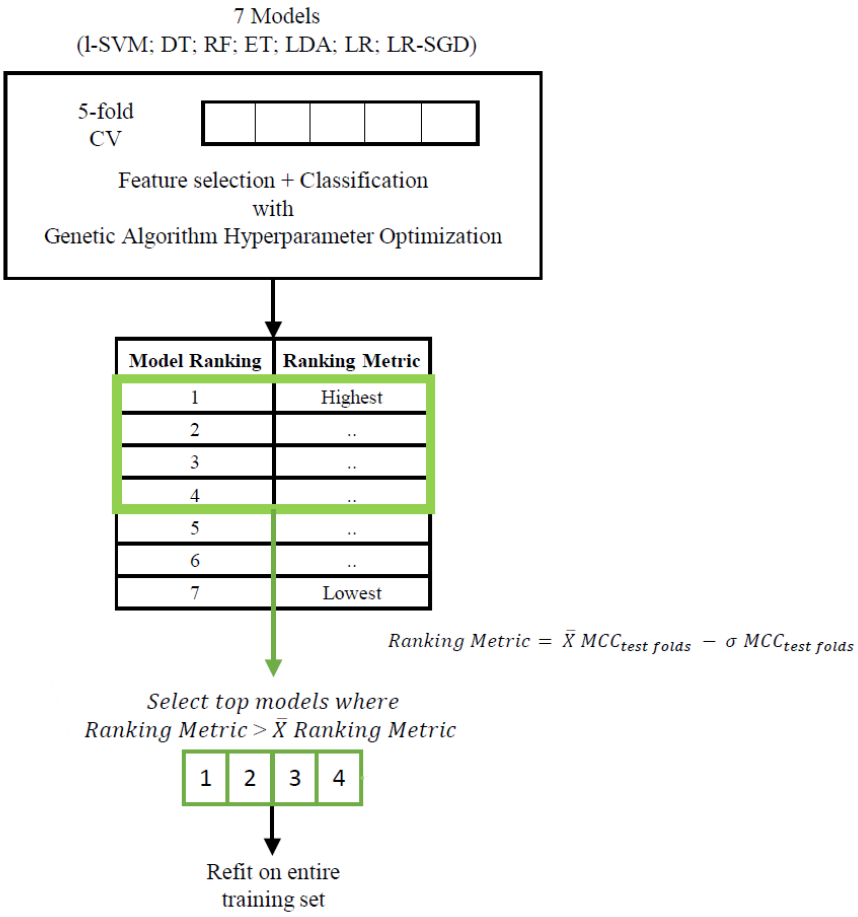


Figure 4.7: Overview of the training of the models

4.4.2 Testing the models

When testing the models on any population, the process is always the same. A dataset containing the data extracted from the MRIs is separated into three binary datasets:

1. A dataset containing only patients classified as Healthy or as AD (1 vs 3)
2. A dataset containing only patients classified as Healthy or as MCI (1 vs 2)
3. A dataset containing only patients classified as MCI or as AD (2 vs 3)

For each binary dataset, the models output a combined probability of each model representing the likelihood of that specific patient having a certain diagnose.

$$\textit{Combined probability} = \overline{\mathbf{Outputprobability}}_{eachmodel} \tag{4.2}$$

where $\overline{\mathbf{Outputprobability}}_{eachmodel}$ is the average probability returned by each model.

In the case of 1 vs 3 and 2 vs 3 the output of the models represent whether a patient has AD or not. The case of 1 vs 2 represents the likelihood of having MCI, as it is possible to see in figure 4.8

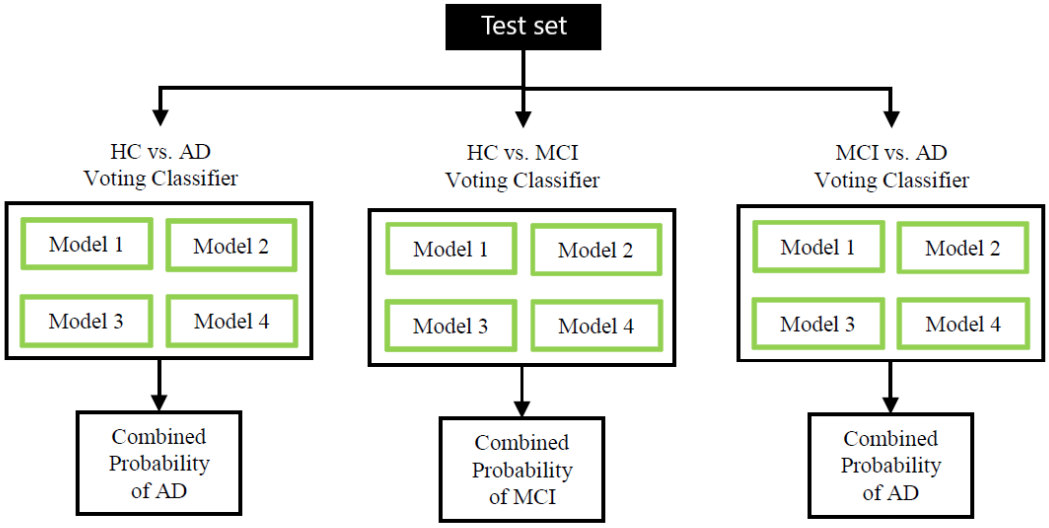


Figure 4.8: Overview of the testing stage of the target predictive models

4.5 External Validation

The goal of this section is to describe the process that took place in order to assess the performance of the models in a real-life situation. For that measure, the models were executed using data from hospitals that are partners of the project. Several hospitals are joining but in the time this thesis was developed only two hospitals, Hospital Vila Franca de Xira and Hospital Fernando Fonseca, were able to supply medical images in the available time span.

4.5.1 Measures in training

In order to assess how well the models are generalizing we considered the use of learning curves and calculation of the bias and variance which allowed understand the following:

1. The variation of performance by varying the number of patients used in the training process;

2. If the models are properly fitted or if they are overfitting or underfitting;

3. If the dataset used in the training and in the validation is representative of the population;

With that said, let us first get into the dataset used to train the models. Such dataset is composed of patients classified as Control, MCI or AD. Figure 4.9 displays the distribution by class and gender, figures D.1, D.2 and D.3, in appendix D, display the distribution by age group of the original population for class Control, MCI and AD, respectively.

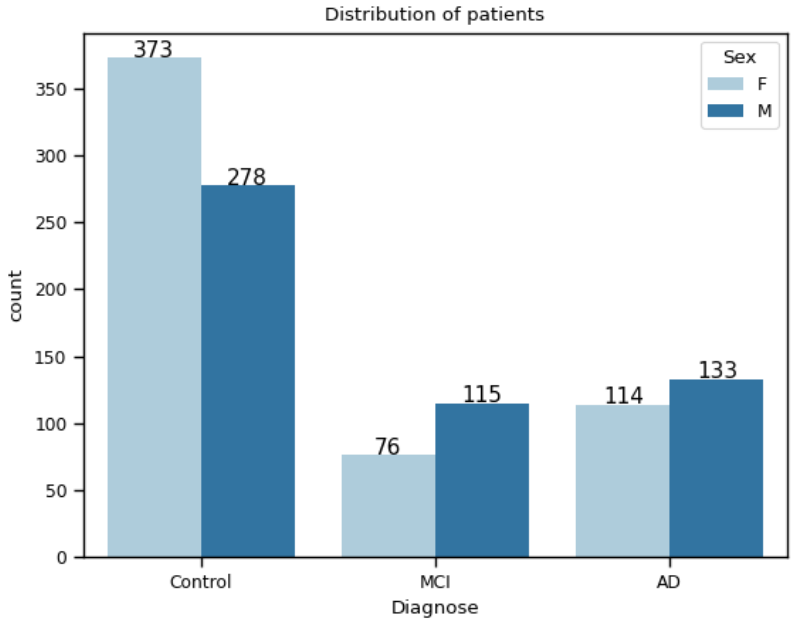


Figure 4.9: Distribution by class and gender of the original population

As previously seen in Section 4.4, where the models are described, there are seven models for each one of the three different scenarios. The dataset that contains the patients from Figure 4.9 represents the original population which is then split so that 70% of the available data is used in the training leaving the other 30% for testing. These models were trained with patients who were classified with either one of the two classes in a scenario, i.e.:

1. **Control vs MCI**: Consists on a dataset of patients classified by the hospitals as Control or MCI
2. **MCI vs AD**: Consists on a dataset of patients classified by the hospitals as MCI or AD
3. **Control vs AD**: Consists on a dataset of patients classified by the hospitals as Control or AD.

Each dataset of each scenario is then divided in order to plot the learning curves, i.e. 80% is used to plot the training error curve and the or 20% is used for the validation curve so, in order to assess the different learning rates of the models by the number of observations (patients), the learning curves were plotted for the group sizes presented in table 4.1. It is important to note that the maximum value in each group size represents the entire 80% mentioned before:

Model	Group sizes
Control vs MCI	[1, 40, 80, 120, 160, 200, 240, 280, 312]
MCI vs AD	[1, 40, 80, 120, 160, 200, 240, 280, 296]
Control vs AD	[1, 40, 80, 120, 160, 200, 240, 280, 303]

Table 4.1: Different dataset sizes for the learning curves plotting.

In addition, in order to complement the analysis of the learning curves, the bias-variance trade-off was another metric that was implemented. Such metric allows us to get a better insight on whether or not the model is overfitting or underfitting so, for that measure, the average expected loss, average bias, and the average variance were calculated for each one of the models.

After all the models were trained, the models were compared with each other in order to analyze the variance thus, allowing us to understand if there is any difference overall between the models. To evaluate such difference, it was used the One-way ANOVA instead of a t-test since it is a parametric test that tests for statistically significant differences between three or more models whereas a t-test allows for just two. The data analyzed by the One-way ANOVA were the values from the accuracy from each one of the five folds of each model, as mentioned in section 4.4.1.

It is important to know that before running the One-way ANOVA, there are some assumptions that were verified as the One-way ANOVA depends on such dependencies to work [51], which are:

1. The distribution of the values from all five folds in each model must be normal, a condition verified using the Shapiro-Wilk test.

2. All models must be independent of each other.

3. All models must have equal variances.

Despite analyzing if there is a statistical significant difference between the predictive models, it is also necessary to see which models differ or not from other models. That way it is possible to compare models in pairs by conducting a Post-Hoc testing using the Bonferroni correction.

4.5.2 Testing the models on a Portuguese population

After the analysis performed in the training dataset , the next stage of the work is to run the models on a Portuguese population. As mentioned before, hospitals provided MRIs for their patients, with the accompanying diagnostics. Such hospitals were Hospital Vila Franca de Xira (HVFX) and Hospital Fernando Fonseca (HFF).

Target population

The target population is composed of HVFX and HFF patients. Figure 4.10 displays the distribution of the population by gender and diagnosis and as it is possible to see there are very few patients diagnosed with AD. In contrast, there is a great number of Control patients. Such imbalance can be explained by two factors:

1. When providing the images, the hospitals prepared patients with dementia and not AD, exclusively so, the patients with AD are a fraction of the whole that is patients with Dementia.

2. When pre-processing the images, the protocols of acquisition of the images (ex: MP-Rage, SGPR, Sag) did not match any protocol accepted by the models.

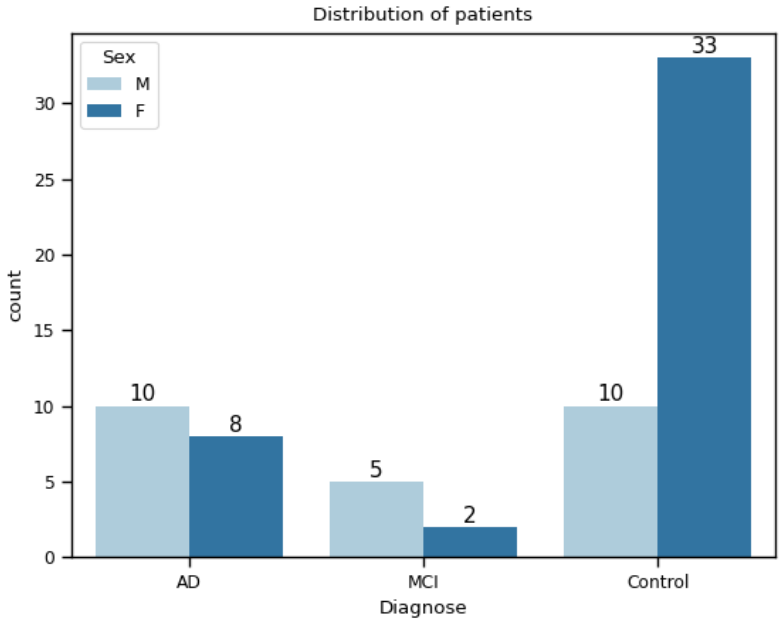


Figure 4.10: Distribution of the target population by gender and diagnose

Regarding the age group of the population, figures D.4, D.5, D.6 display the distribution of the population for each class of interest by age group and, as expected, patients diagnosed as AD or MCI range from 55 years old onwards whereas patients diagnosed as Control range from 25 years old onwards which, once again is expected since AD is a disease that occurs, mostly, at later stages of life.

After getting a general view of the population, it is now time to present the measures that are implemented in order to interpret the performance of the models on said population.

The first evaluation step to be applied is the confusion matrix analysis since it allows us to have a generalized and summarized view of the performance of the models for the multiclass problem. Having calculated the confusion matrix is then possible to get the values of the precision, sensitivity, and specificity, as well as the balanced accuracy of the model. Although accuracy is calculated too, relying just on such metric may be considerably misleading when handling an imbalanced dataset. The confusion matrix and the balanced accuracy can address that as they account for both the positive and negative predicted classes without misleading performance summaries in the presence of imbalanced data.

After having calculated the specificity and the sensitivity, the Area under the ROC (receiver operating characteristic) curve is suggested to assess how well the model is capable of distinguishing between HC, MCI, and AD. With precision and sensitivity calculated as well, the precision-recall curve is also suggested.

Chapter 5

Results and Discussion

5.1 Data anonymization

The developed script receives, as mentioned before, DICOM files that contain confidential information from the patient as seen in figure 4.2. After anonymization the script produces 3 files and the anonymized image itself.

The first produced file, Keys.csv, contains the original identifier from the patient and the new identifier generated by the script. This way, once there is a new classification for the patient, the hospital can re-identify the patient.

The second generated file, PhysicianName.csv, stores the real name of the physician that performed the exam as well as its new ID.

Lastly, AccessionNumber.csv is the last file that is produced, storing the accession number of the exams, which is basically the ID of a specific exam, as well as the new ID created by the script.

All these files are only in the possession of the hospital/clinic so, IBEB has no knowledge of the content of such files.

The images are the most important part of the anonymization and as it is possible to see, figures 5.1 and 5.2 display a frame from an MRI exam from an ADNI patient. Figure 5.1 presents information regarding the patient that must be anonymized. Such information is the patient_ID (009_S_1030), patient_name (009_S_1030), birth date (26/05/1939) and others.

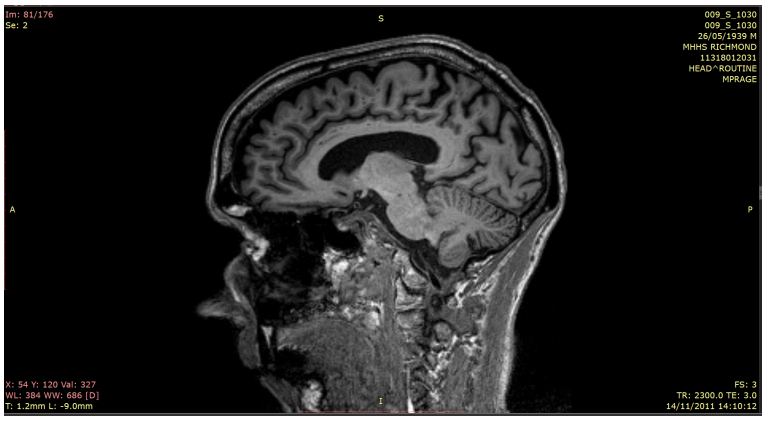


Figure 5.1: MRI image before anonymization.

The anonymization of that image consisted on the deletion of the tags regarding the PHI tags and the result of such process can be seen in figure 5.2.

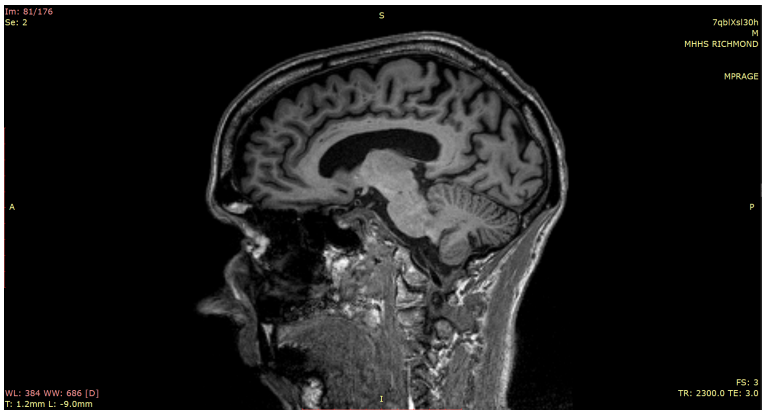


Figure 5.2: MRI image after anonymization.

5.2 Database

The database schema developed is presented in Figure B.4. It consists on a set of tables as follows:

Table Name	Column Name	Content Meaning	Data Type
Patients_Info	Patient_ID Sex Age_First_Visit Preliminary_Diagnosis Hospital	ID of the Patient Gender of the Patient Age of the patient at the first visit at the hospital/clinic First diagnose given to the patient Source of patient	varchar(255) varchar(255) Numeric varchar(255) varchar(255)
All_Exams	Exams_ID Patient_ID Scan_Date	ID of the exam ID of the Patient Date of the exam	varchar(255) varchar(255) Date
Exams_Info	Exams_ID Tag Value	ID of the exam ID of the dicom tag Content of the dicom tag	varchar(255) varchar(255) varchar(255)
Dicom_Tags	Tag Tag_Name Value_Representation Retired	Tag that identifies the attribute Name of the dicom tag Describes the data type and format of Tag value Tells if tag is still being used or not	varchar(255) varchar(255) varchar(255) boolean
Diagnosis_History	Patient_ID Date Diagnosis	ID of the Patient Date of the diagnosis Diagnosis given to the patient	varchar(255) Date varchar(255)
User	Username Email Password Admin	username User's email User's password Tells if user is a admin or not	varchar(15) varchar(50) varchar() boolean
ADNI_Control	Several Variables	Attributes regarding non imaging exams from ADNI	Varchar(255)/ Boolean/ Numeric
PD_Control	Several Variables	Attributes regarding non imaging exams from Parkinson's disease Markers Initiave	Varchar(255)/ Boolean/ Numeric

Table 5.1: Database tables

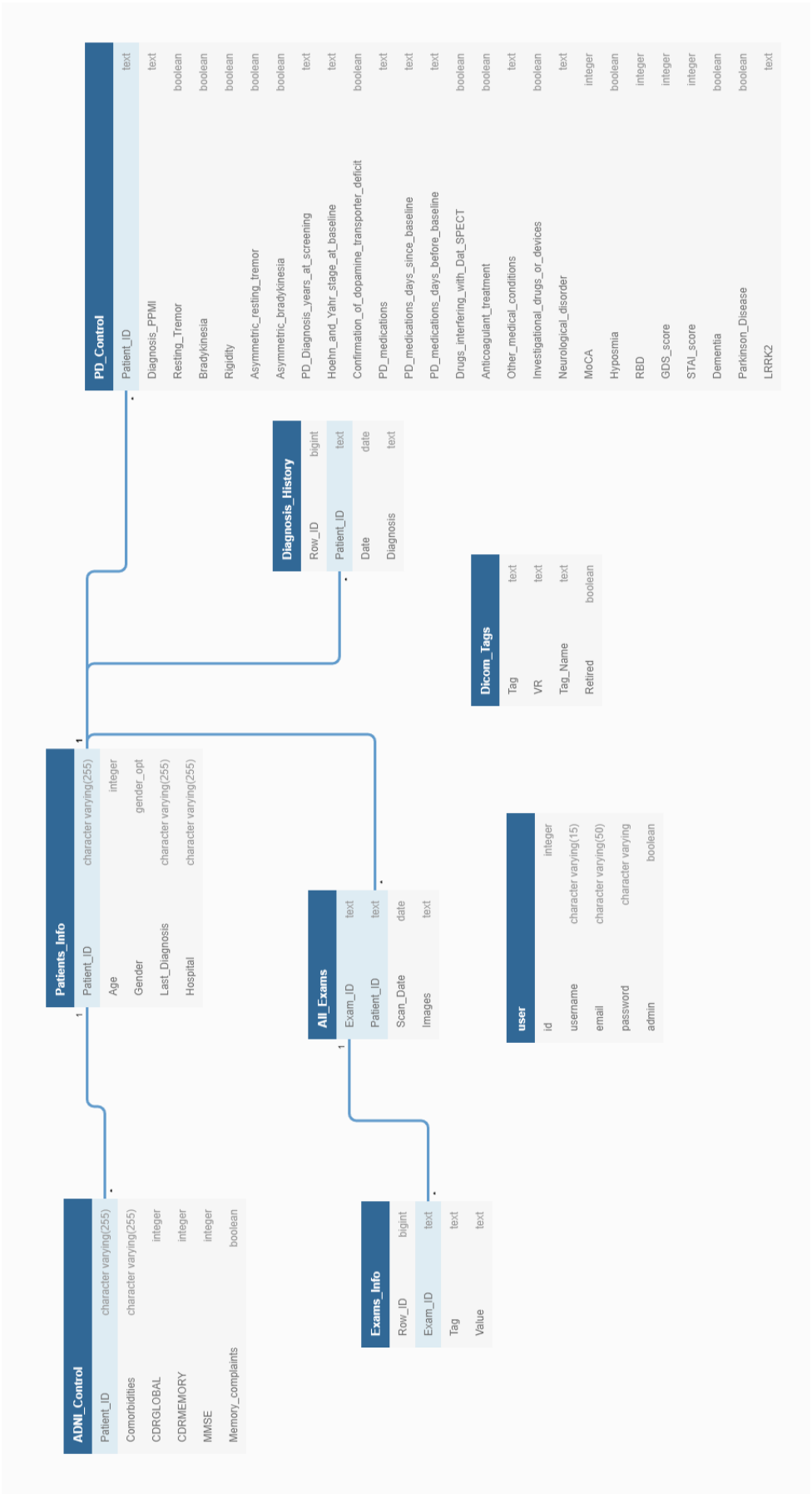


Figure 5.3: General view of the relational database model

A set of questions must be answered by the database. Appendix provides a list of the most important/frequent queries. For illustrative purposes we consider one of said queries:

1. What is the total number of patients with each diagnose for the different age groups?.

   Query:

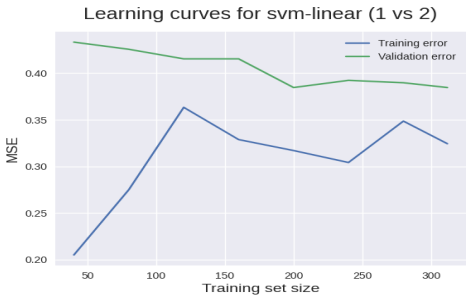
```
SELECT
    "Last_Diagnosis",
    "Age"
    Count("Age") as Number_of_patients
FROM
    public."Patients_Info"
GROUP BY
    ("Last_Diagnosis", "Age");
```

5.3 Statistical Validation

This section has the purpose of presenting and discussing the results obtained on the original population of the ADNI initiative and the target Portuguese population. Section 5.3.1 presents the results obtained for the learning curves, bias-variance trade-off and the ANOVA tests implemented in the models as soon as they were trained. Section 5.3.2 presents the results of the validation performed on the models. Such validation was performed on the original data available and on the the target data which represents the the patients from Hospital Vila Franca de Xira and Hospital Fernando Fonseca.

5.3.1 Validation using the original (heterogeneous) population

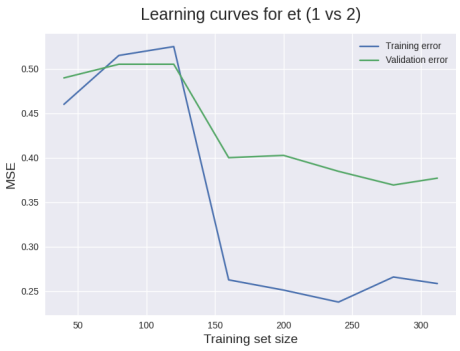
In order to assess the generalization ability of models, i.e how they change in terms of performance over different population sizes as well as seeing if any model is underfitting or overfitting, Figures 5.4a, 5.5a, 5.6a show the plotted learning curves for each one of the seven models in each one of the three possible scenarios (Control vs MCI, Control vs AD and MCI and AD). The plotted figures display the mean square error for both the validation set and the training set.



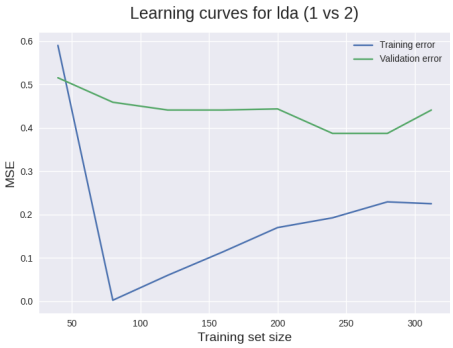
(a)

(b)

(c)

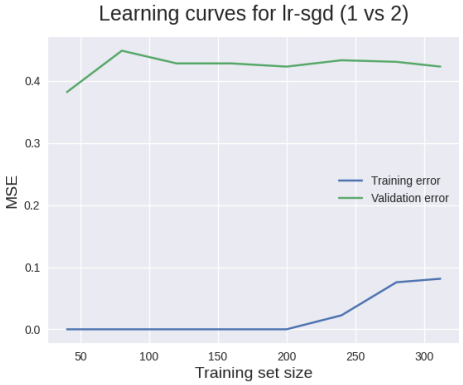


(d)



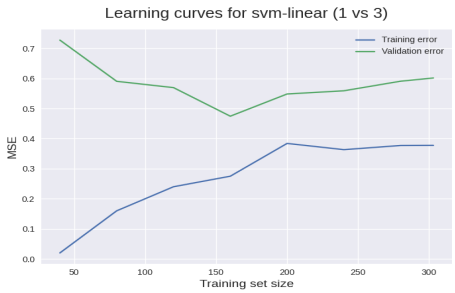
(e)

(f)



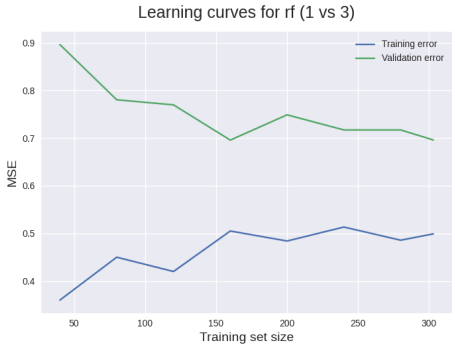
(g)

Figure 5.4: Learning curves in the Control vs MCI scenario for SVM-Linear (a), Decision Trees (b), Random Forests (c), Extra Trees (d), Linear discriminant analysis (e), Logistic Regression (f) and Logistic Regression with Stochastic Gradient Descent (g) when training the models.

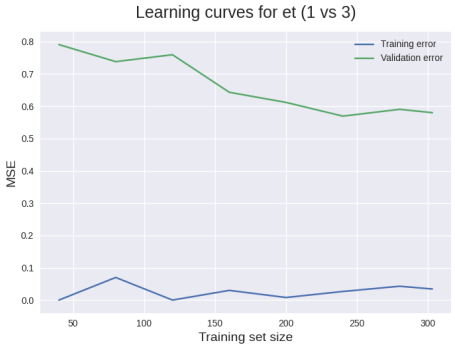


(a)

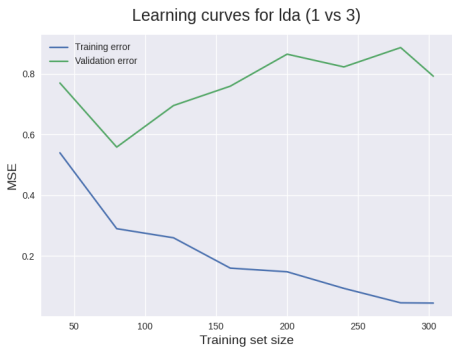
(b)



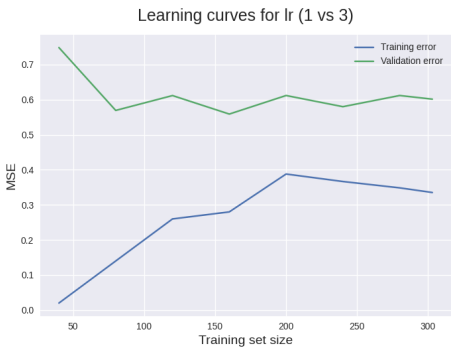
(c)



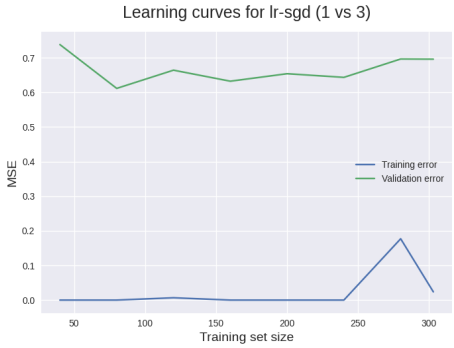
(d)



(e)

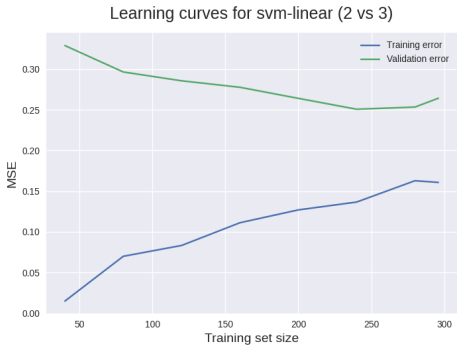


(f)

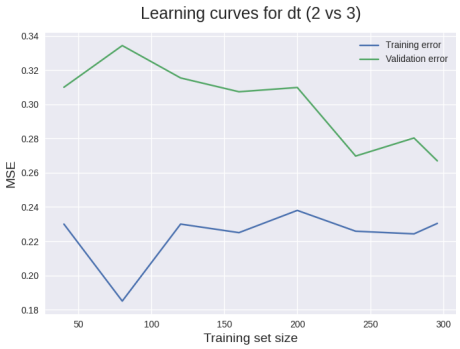


(g)

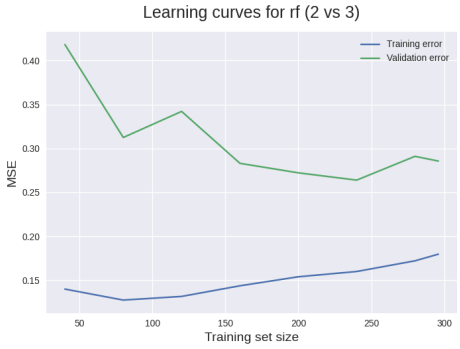
Figure 5.5: Learning curves in the Control vs AD scenario for SVM-Linear (a), Decision Trees (b), Random Forests (c), Extra Trees (d), Linear discriminant analysis (e), Logistic Regression (f) and Logistic Regression with Stochastic Gradient Descent (g) when training the models.



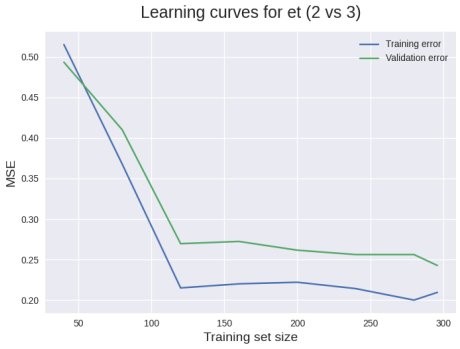
(a)



(b)



(c)



(d)

(e)

(f)

(g)

Figure 5.6: Learning curves in the MCI vs AD scenario for SVM-Linear (a), Decision Trees (b), Random Forests (c), Extra Trees (d), Linear discriminant analysis (e), Logistic Regression (f) and Logistic Regression with Stochastic Gradient Descent (g) when training the models.

Similar to the learning curves, the bias and variance are also calculated in order to complement the analysis, assessing how well-fitted are the targeted models. Tables 5.2, 5.3, 5.4 show the results obtained for each one of the seven models in each one of the three possible scenarios (Control vs MCI, Control vs AD, MCI and AD) of the bias, variance and the expected loss.

	Control vs MCI						
	SVM-Linear	**DT**	**RF**	**ET**	**LR**	**LDA**	**LR-SGD**
Average expected loss	0.376	0.426	0.390	0.339	0.288	0.333	0.200
Average bias	0.294	0.299	0.257	0.199	0.170	0.151	0.060
Average variance	0.081	0.127	0.134	0.141	0.118	0.182	0.140

Table 5.2: Bias and variance for each model in the Control vs MCI scenario.

	Control vs AD						
	SVM-Linear	**DT**	**RF**	**ET**	**LR**	**LDA**	**LR-SGD**
Average expected loss	0.386	0.545	0.465	0.201	0.366	0.352	0.231
Average bias	0.253	0.210	0.313	0.057	0.238	0.086	0.057
Average variance	0.133	0.335	0.152	0.144	0.128	0.267	0.173

Table 5.3: Bias and variance for each model in the Control vs AD scenario

	MCI vs AD						
	SVM-Linear	**DT**	**RF**	**ET**	**LR**	**LDA**	**LR-SGD**
Average expected loss	0.179	0.252	0.180	0.190	0.187	0.193	0.167
Average bias	0.100	0.148	0.091	0.142	0.108	0.086	0.068
Average variance	0.079	0.104	0.089	0.049	0.079	0.106	0.099

Table 5.4: Bias and variance for each model in the MCI vs AD scenario

The analysis of the learning curves 5.4, 5.5, 5.6 and tables 5.2, 5.3, 5.4 allows us to see that for scenario:

1. **Control vs MCI**, decision trees will not be benefit from the increase of instances in the training set since the validation and training error curves have already converged. LR-SGD may be overfitting since the validation error is high wheres as the training error is much lower resulting in a high variance;
2. **Control vs AD**, LDA is overfitting as a result of the decreasing training error and the increasing of the validation error. A case of overfitting may be identified in the extra trees due to the low bias and the relatively higher variance;
3. **MCI vs AD**, models seem to present lower levels of bias and a higher value of variance, with the exception of SVM-linear.

The overall conclusion, is that the models may not generalise as well as they could as seen by the low bias and higher variance, hence the much higher validation error when compared against the training error. The training set sizes do not allow for an extensive analysis of the models so the best recommendation would be to increase the instances available in the training, i.e. gather a higher number of patients which can be used to re-train the models so that the learning curves could show the validation and training error curves converged which it not happening for the most cases.

Comparing the models with ANOVA

After the aforementioned metrics were calculated and the training process was concluded, it was also necessary to compare the models with each other and see if they are equal in any way. Since we need our data to follow a normal distribution and the variance must be the same for all the data [51], table 5.5 displays the results from the Saphiro-Wilk test where it compares the balanced accuracy from the folds of each model. The results prove that the data follows as normal distribution as the p-value is above 0.05 in all cases for each one of the model.

Model	***svm-linear***	***dt***	***rf***	***et***	***lda***	***lr***	***lr-sgd***
p-value (Control vs MCI)	0.109	0.637	0.557	0.967	0.669	0.794	0.515
p-value (MCI vs AD)	0.771	0.062	0.763	0.437	0.592	0.147	0.196
p-value (Control vs AD)	0.414	0.399	0.071	0.071	0.918	0.348	0.155

Table 5.5: P-values of each model for each one of the three scenarios

Regarding the homogeneity of variance table 5.6 shows the p-values obtained after comparing the models in each one of the three scenarios under the Levene's test. As it is possible to see, the Levene's Test [52] of homogeneity of variances is not significant (p-values $>$ 0.05) thus, concluding there is no statistical difference in the variability of the models within a scenario.

	Control vs MCI	***Control vs AD***	***MCI vs AD***
P-values	0.969	0.99	0.834

Table 5.6: Levene's test for each scenario

After verifying the assumptions above, a Post-Hoc Test [53] was performed to see which models significantly differ from each others. The Post-Hoc test with the Bonferroni correction returned false for all the pairs of models comapred which means that, no model differs significantly from other models.

5.3.2 Generalization analysis in a Portuguese population

When running the models for the target population (Hospital Vila Franca de Xira + Hospital Fernando Fonseca), the confusion matrix in Figure 5.7a is obtained. Table 5.7 complements the confusion matrix since it allows us to know the values of precision, recall/sensitivity, F1-score, and the number of patients that support such calculus.

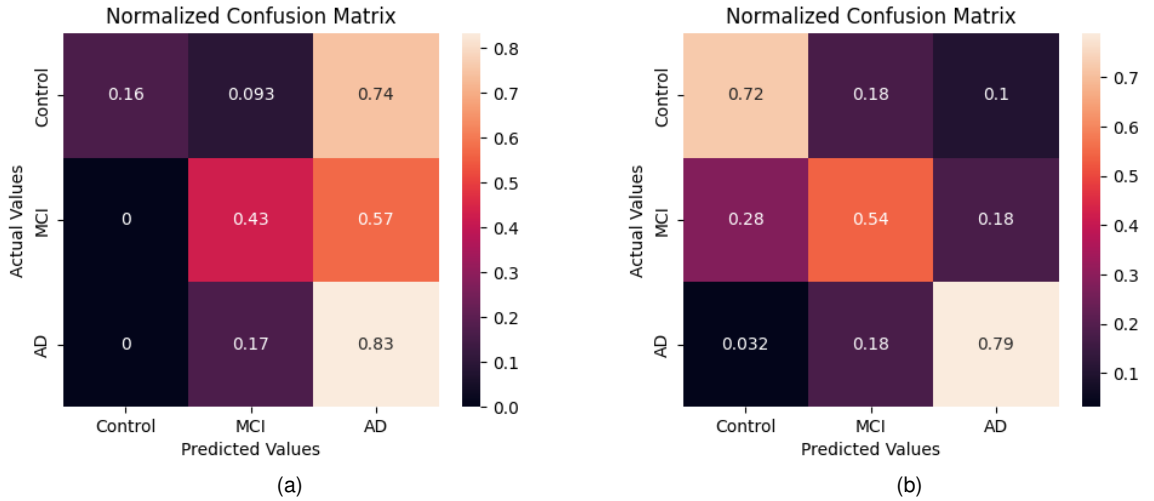


Figure 5.7: Confusion matrix of the target population 5.7a and the original population 5.7b

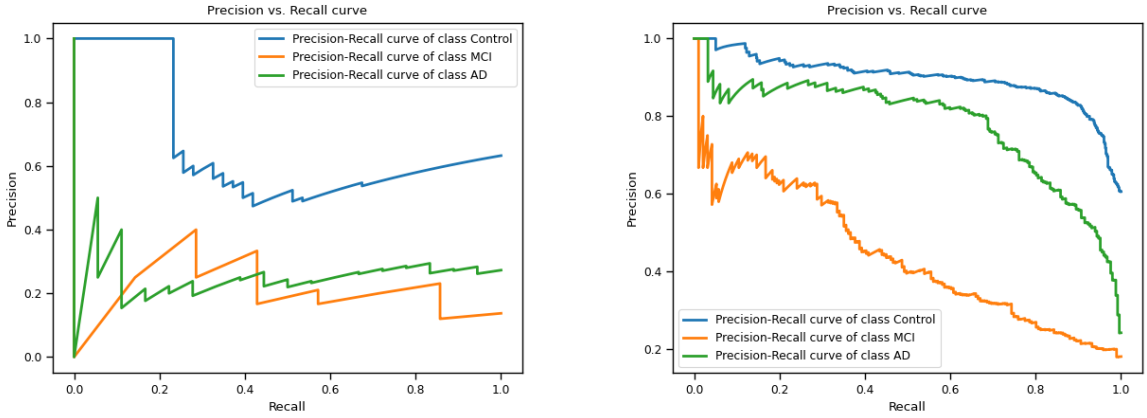
	Target Population				Original Population			
	Precision	**Recall**	**F1-score**	**Support**	**Precision**	**Recall**	**F1-score**	**Support**
Control	1.00	0.16	0.28	43	0.88	0.72	0.79	651
MCI	0.30	0.43	0.35	7	0.39	0.54	0.45	191
AD	0.29	0.83	0.43	18	0.66	0.79	0.72	247

Table 5.7: Predictive accuracy for Portuguese population

Starting with the class Control, it is clear that the models do not classify any patients as Control when they should not be classified as such hence the precision value of 1 (100%) which in other words mean that the fraction of instances correctly predicted as Control is 100% out of the total classified instances as Control. On the other hand, the value of the recall is only 0.16 which means that out of the 43 patients, the models might not have wrongly classified AD or MCI patients as Control but, the low value of recall means that 84% (36 patients) of the Control patients were classified as either MCI or AD patients.

In the case of the MCI patients, the precision value decreases substantially from 1 to 0.30 but on the other hand, the recall value increased from 0.16 to 0.43. One might say that these values are preferable when looking at f1-score which is higher. Out of all the patients classified as MCI, the models lacked the ability to accurately find all the MCI patients since 57% of the MCI patients were classified as AD. Such low values may be explained by the low support value of patients (only 7 patients in the entire target population).

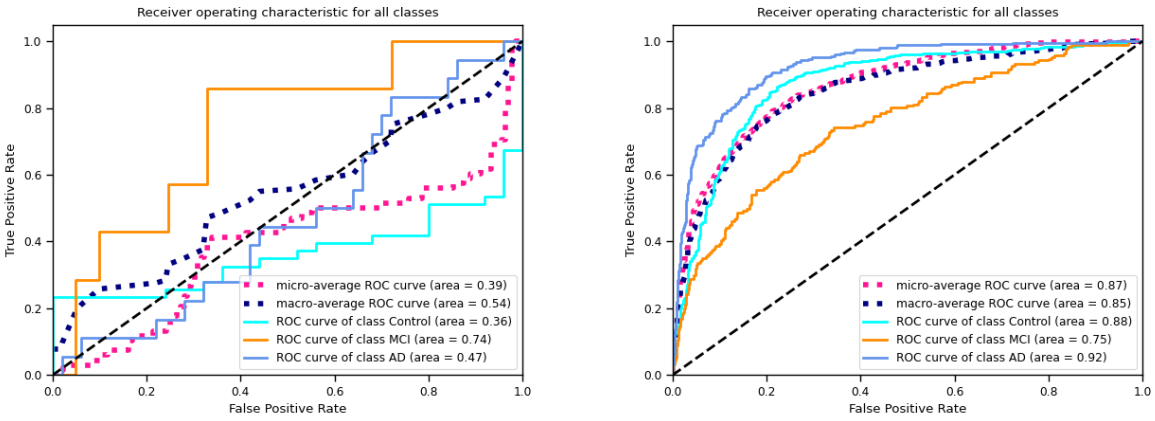
Lastly, looking at the AD patients, the precision value is 0.29, which represents that out of all the patients classified as AD only 29% of those were correctly classified as AD. In contrast, 83% of the patients with AD were correctly classified as AD.



(a) (b)

Figure 5.8: Precision Recall curve obtained in the target population 5.8a and the original population 5.8b

The ROC curves displayed in figure 5.9a represent the trade-off between sensitivity and specificity. Such curves are useful since they do not rely on the distribution of classes which comes in handy considering the number of control patients is not balanced with the number of AD patients (43 control to 18 AD patients) and allows for the better interpretation of the MCI class (7 patients).



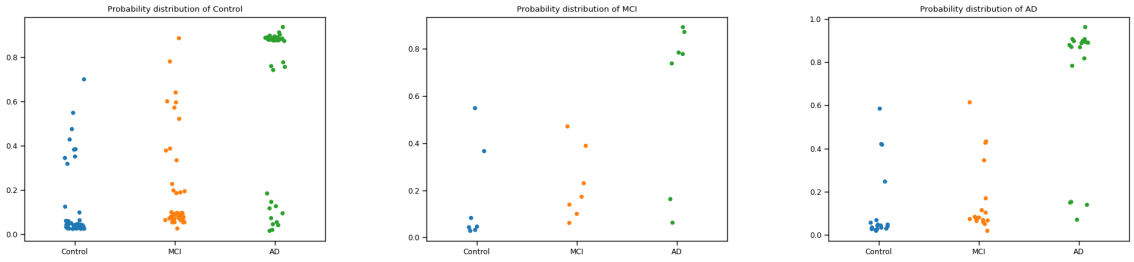
(a) (b)

Figure 5.9: ROC curves obtained in the target population obtained in the target population 5.9a and the original population 5.9b

The models can be interpreted by comparing their performance against a baseline which is the FPR = TPR diagonal that represents the expected values a random classifier would return. The models' performance is considered low since the curves are closer to the 45 degrees diagonal when they should be closer to the top-left corner of the graph as it is the case on the original population.

In addition, to obtain a better view of the models output, figure 5.10 displays the distribution of probabilities for the patients diagnosed by the hospitals as Control, MCI and AD. The main goal of said figure is to show that, for instance, Control patients are classified as AD as proved by the high probability in the AD column, hence the lack of ability to predict control patients. The same event occurs for the MCI patients as the class with higher probabilities is AD instead of MCI. As mentioned before, 83% of the patients classified as AD were in fact AD patients. The predicted value is given with a high level of

confidence hence the low values of control and MCI probability in 5.10c.



(a) (b) (c)

Figure 5.10: Distribution of calculated probabilities for patients diagnosed by the hospitals as Control 5.10a, MCI 5.10b or AD 5.10c

Overall, it is possible to see that the models cannot generalise well for the target population as they generalize for the original data. Such disparity in results may be explained by the relatively small size of training set in which the models were trained or even by the fact that the exams used when validating the models presented new protocols of acquisition of MRIs.

Chapter 6

Conclusion

The work presented in this dissertation focuses on a specific dementia disorder, Alzheimer's disease, which accounts for two-thirds of all dementia cases. The fact that diagnosing a patient with Alzheimer's disease at an early stage of development is not always straightforward does not rule out the possibility of more powerful diagnostic tools being developed in the future. In reality, by utilizing various sorts of technologies, it is feasible to diagnose individuals with Alzheimer's disease with high precision. The tools used nowadays are based on the examination of the patient's and their families' medical histories, after which neuropsychological tests can be performed to assess cognitive function but with the downside that such tests may take time and might be performed at a later stage of the disease.

Machine learning approaches are used in this work to construct predictive models capable of delivering an early and accurate diagnosis of Alzheimer's disease or even a preclinical stage of cognitive impairment preceding Alzheimer's at a later period in life. The most significant challenges of such models is the necessity to ensure their interpretability and ability to handle complex data, as well as their ability to generalize on external data since the data collected to feed the classification models is large and heterogeneous in nature. The specified data consists on Medical Resonance Imaging records, which contain not only medical information but also demographic information about the patient. The data collected from the hospitals needed to be anonymized before receiving said images in order to maintain the anonymity of the patients and the medical professionals. Furthermore, it was necessary to build a centralized solution capable of storing the data so it could be later used in the study. This dissertation aimed to validate models under a heterogeneous population to ensure adequate representation of the Portuguese population and guarantee sufficient generalization capability of the models. In face of all the requirements, a relational database was developed in order to store the content of the MRIs received from the hospital. Prior to the reception of said data, a script was developed so that the MRIs could be anonymized and later sent. The existence of a database resulted in the development of Graphic User Interface capable of manipulating the database by allowing the user to insert new data or view data and statistics from the database content. In addiction, a validation of the predictive models was performed, being the latter the main focus of this dissertation. To assess the generalization ability of the models, these are tested on a target population consisting of patients from Hospital Vila Franca de Xira and Hospital Fernando Fonseca. Before running the models on the target population, learning curves are plotted which together with the bias and variance calculus allow to understand whether or not the models are adequately fit on the data. ANOVA and a Post-Hoc test are used in order to compare the models with each other and see if they were equal in any way. To figure how well the models were generalizing for the target population, commonly used measures were calculated in order to extract statistics on the capabilities of the models.

The learning curves showed that the models are not yet at a point of maturity. Both the learning

curves and the bias-and-variance calculus allowed to understand that some models could be facing an underfitting or overfitting problem when handling Control vs MCI patients, as is the case of the models with Logistic Regression or with Linear Discriminant Analysis due to high values of bias. In addition, the learning curves showed, in the case of Linear Discriminant Analysis, a decreasing training error and an increasing validation error for an increased data size. The Post-Hoc test showed that all the pairs of models compared presented no significant difference in the variability of the models within each scenario. Regarding the results on the target population, the models showed that they lack the ability to generalize well on the new population as they did on the original one. For class Control, 100% was achieved on precision whereas in the case of Recall or F1-score only 16% and 28%, respectively, was achieved. Class MCI had slightly different results with 30% of precision, 43% of recall and 35% of F1-score while the AD class presented a 29% of precision, 83% of recall and 43% of F1-score. Furthermore, the Area Under the Receiver Operating Characteristics for class Control showed an area of 36%, followed by MCI area of 74% and AD with 0.47%. Overall, the results showed that the models do not have the desired ability to generalize well for a new population. Although the results for the AD class were better and no false negative were returned for the control class, i.e. no patient with Alzheimer's disease was classified as control, the models did not perform well on the population.

The main hypothesized reason for this inability to generalize well is the low volume of available patients used in the training of the models. One would recommend adding more patients to the training process so that the models could achieve optimal performance in training. The learning curves showed the training and validation error curve did not converge due to the low volume of instances so, by adding more patients this problem could be solved. Another reason for the low generalization capacity is considered to be the different image acquisition protocols of the MRIs in the target population.

6.1 Future Work

Despite the relevance of the produced results from the targeted models, there are some measures that could be implemented in order to improve the overall performance of the models. First, the extension of the predictive models to another population since new populations result in different data that might correlate better with the models' requirements, so that models could be retrained under a more heterogeneous populations and, consequently, generalize better on a new target population. In addition, it would be interesting to see alternative supervised classification principles to assess whether the performance of the models. Third, the proposed data consolidation and external validation principles can be considered to expand the scope of the work in order to handle new neurological diseases. Finally, regarding the database and the GUI, it would be interesting to deploy the database onto a remote server so that several entities could access it and easily insert new data with the necessary guarantees of security, privacy and usability.

Bibliography

[1] "Understanding the bias-variance tradeoff — by seema singh — towards data science." https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229. (Accessed on 01/07/2021).

[2] "The intuition behind bias and variance — by seth mottaghinejad — towards data science." https://towardsdatascience.com/bias-and-variance-but-what-are-they-really-ac539817e171. (Accessed on 10/30/2021).

[3] C. Perlich, "Learning curves in machine learning.," 2010.

[4] M. J. Zaki and W. Meira, *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[5] M. Sayadi, M. Zibaeenezhad, and S. M. Taghi Ayatollahi, "Simple prediction of type 2 diabetes mellitus via decision tree modeling," *International Cardiovascular Research Journal*, vol. 11, no. 2, pp. 71–76, 2017.

[6] D. Speelman, "Logistic regression a confirmatory technique for comparisons in corpus linguistics."

[7] C. R. Jack Jr, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski, "Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade," *The Lancet Neurology*, vol. 9, no. 1, pp. 119–128, 2010.

[8] C. Adelina, "The costs of dementia: advocacy, media and stigma.," *Alzheimer's Disease International: World Alzheimer Report 2019*, pp. 100–101, 2019.

[9] R. L. Nussbaum and C. E. Ellis, "Alzheimer's disease and parkinson's disease," *New england journal of medicine*, vol. 348, no. 14, pp. 1356–1364, 2003.

[10] L. Mucke, "Alzheimer's disease," *Nature*, vol. 461, no. 7266, pp. 895–897, 2009.

[11] H. Ung, J. E. Brown, K. A. Johnson, J. Younger, J. Hush, and S. Mackey, "Multivariate classification of structural mri data detects chronic low back pain," *Cerebral cortex*, vol. 24, no. 4, pp. 1037–1044, 2014.

[12] D. Prayer, P. C. Brugger, and L. Prayer, "Fetal mri: techniques and protocols," *Pediatric radiology*, vol. 34, no. 9, pp. 685–693, 2004.

[13] R. C. Petersen, R. Doody, A. Kurz, R. C. Mohs, J. C. Morris, P. V. Rabins, K. Ritchie, M. Rossor, L. Thal, and B. Winblad, "Current concepts in mild cognitive impairment," *Archives of neurology*, vol. 58, no. 12, pp. 1985–1992, 2001.

[14] A. J. Saykin, L. Shen, X. Yao, S. Kim, K. Nho, S. L. Risacher, V. K. Ramanan, T. M. Foroud, K. M. Faber, N. Sarwar, *et al.*, "Genetic studies of quantitative mci and ad phenotypes in adni: Progress, opportunities, and plans," *Alzheimer's & Dementia*, vol. 11, no. 7, pp. 792–814, 2015.

[15] R. J. Killiany, T. Gomez-Isla, M. Moss, R. Kikinis, T. Sandor, F. Jolesz, R. Tanzi, K. Jones, B. T. Hyman, and M. S. Albert, "Use of structural magnetic resonance imaging to predict who will get alzheimer's disease," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 47, no. 4, pp. 430–439, 2000.

[16] R. A. McArthur, *Translational neuroimaging: tools for CNS drug discovery, development and treatment*. Academic Press, 2012.

[17] N. K. Logothetis, "What we can do and what we cannot do with fmri," *Nature*, vol. 453, no. 7197, pp. 869–878, 2008.

[18] A. J. van der Kouwe, T. Benner, D. H. Salat, and B. Fischl, "Brain morphometry with multiecho mprage," *Neuroimage*, vol. 40, no. 2, pp. 559–569, 2008.

[19] J. Wang, L. He, H. Zheng, and Z. L. Lu, "Optimizing the Magnetization-Prepared Rapid Gradient-Echo (MP-RAGE) sequence," *PLoS ONE*, vol. 9, no. 5, 2014.

[20] N. Patronas, N. Bulakbasi, C. A. Stratakis, A. Lafferty, E. H. Oldfield, J. Doppman, and L. K. Nieman, "Spoiled gradient recalled acquisition in the steady state technique is superior to conventional postcontrast spin echo technique for magnetic resonance imaging detection of adrenocorticotropin-secreting pituitary tumors," *Journal of Clinical Endocrinology and Metabolism*, vol. 88, no. 4, pp. 1565–1569, 2003.

[21] M. Mustra, K. Delac, and M. Grgic, "Overview of the dicom standard," in *2008 50th International Symposium ELMAR*, vol. 1, pp. 39–44, IEEE, 2008.

[22] R. C. Petersen, P. Aisen, L. A. Beckett, M. Donohue, A. Gamst, D. J. Harvey, C. Jack, W. Jagust, L. Shaw, A. Toga, *et al.*, "Alzheimer's disease neuroimaging initiative (adni): clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.

[23] N. H. Strickland, "Pacs (picture archiving and communication systems): filmless radiology," *Archives of disease in childhood*, vol. 83, no. 1, pp. 82–86, 2000.

[24] E. Vargiu and M. Urru, "Exploiting web scraping in a collaborative filtering-based approach to web advertising.," *Artif. Intell. Research*, vol. 2, no. 1, pp. 44–54, 2013.

[25] C. for Disease Control, Prevention, *et al.*, "Hipaa privacy rule and public health. guidance from cdc and the us department of health and human services," *MMWR: Morbidity and mortality weekly report*, vol. 52, no. Suppl 1, pp. 1–17, 2003.

[26] K. Strimbu and J. A. Tavel, "What are biomarkers?," *Current Opinion in HIV and AIDS*, vol. 5, no. 6, p. 463, 2010.

[27] N. Cristianini and E. Ricci, *Support Vector Machines*, pp. 928–932. Boston, MA: Springer US, 2008.

[28] R. Amami, D. B. Ayed, and N. Ellouze, "Practical selection of svm supervised parameters with different feature representations for vowel recognition," *arXiv preprint arXiv:1507.06020*, 2015.

[29] L. Breiman, "Random forests," 2001.

[30] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: A classification and regression tool for compound classification and qsar modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, pp. 1947–1958, 11 2003.

[31] A. M. Martinez and A. C. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 228–233, 2 2001.

[32] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[33] V. I. Volosnikov, V. V. Korkhov, A. O. Vorontsov, K. V. Gribkov, A. B. Degtyarev, A. V. Bogdanov, N. M. Zalutskaya, N. G. Neznanov, and N. I. Ananyeva, "Data consolidation and analysis system for brain research," *CEUR Workshop Proceedings*, vol. 2267, no. Grid, pp. 388–392, 2018.

[34] S. Nugawela, *Data warehousing model for integrating fragmented electronic health records from disparate and heterogeneous clinical data stores*. PhD thesis, Queensland University of Technology, 2013.

[35] W. Newhauser, T. Jones, S. Swerdloff, W. Newhauser, M. Cilia, R. Carver, A. Halloran, and R. Zhang, "Anonymization of dicom electronic medical records for radiation therapy," *Computers in biology and medicine*, vol. 53, pp. 134–140, 2014.

[36] A. Act, "Health insurance portability and accountability act of 1996," *Public law*, vol. 104, p. 191, 1996.

[37] G. D. P. Regulation, "General data protection regulation (gdpr)," *Intersoft Consulting, Accessed in October 24*, vol. 1, 2018.

[38] Y. Zhu, P. Singh, K. Siddiqui, and M. Gillam, "An automatic system to detect and extract texts in medical images for de-identification," in *Medical Imaging 2010: Advanced PACS-based Imaging Informatics and Therapeutic Applications*, vol. 7628, p. 762803, International Society for Optics and Photonics, 2010.

[39] M. Onken, J. Riesmeier, M. Engel, A. Yabanci, B. Zabel, and S. Després, "Reversible anonymization of dicom images using automatically generated policies.," in *MIE*, pp. 861–865, 2009.

[40] D. Abouakil, J. Heurix, and T. Neubauer, "Data models for the pseudonymization of dicom data," in *2011 44th Hawaii International Conference on System Sciences*, pp. 1–11, 2011.

[41] D. R. González, T. Carpenter, J. I. van Hemert, and J. Wardlaw, "An open source toolkit for medical imaging de-identification," *European radiology*, vol. 20, no. 8, pp. 1896–1904, 2010.

[42] M. Bocchetta, S. Galluzzi, P. G. Kehoe, E. Aguera, R. Bernabei, R. Bullock, M. Ceccaldi, J.-F. Dartigues, A. De Mendonca, M. Didic, *et al.*, "The use of biomarkers for the etiologic diagnosis of mci in europe: An eadc survey," *Alzheimer's & Dementia*, vol. 11, no. 2, pp. 195–206, 2015.

[43] S. Qiu, P. S. Joshi, M. I. Miller, C. Xue, X. Zhou, C. Karjadi, G. H. Chang, A. S. Joshi, B. Dwyer, S. Zhu, M. Kaku, Y. Zhou, Y. J. Alderazi, A. Swaminathan, S. Kedar, M.-H. Saint-Hilaire, S. H. Auerbach, J. Yuan, E. A. Sartor, R. Au, and V. B. Kolachalama, "Development and validation of an interpretable deep learning framework for Alzheimer's disease classification," *Brain*, vol. 143, pp. 1920–1933, 05 2020.

[44] S. Bleeker, H. Moll, E. Steyerberg, A. Donders, G. Derksen-Lubsen, D. Grobbee, and K. Moons, "External validation is necessary in prediction research:: A clinical example," *Journal of clinical epidemiology*, vol. 56, no. 9, pp. 826–832, 2003.

[45] G. C. Siontis, I. Tzoulaki, P. J. Castaldi, and J. P. Ioannidis, "External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination," *Journal of Clinical Epidemiology*, vol. 68, no. 1, pp. 25 – 34, 2015.

[46] K. G. Moons, A. P. Kengne, D. E. Grobbee, P. Royston, Y. Vergouwe, D. G. Altman, and M. Woodward, "Risk prediction models: Ii. external validation, model updating, and impact assessment," *Heart*, vol. 98, no. 9, pp. 691–698, 2012.

[47] E. W. Steyerberg and F. E. Harrell, "Prediction models need appropriate internal, internal–external, and external validation," *Journal of clinical epidemiology*, vol. 69, pp. 245–247, 2016.

[48] K. Y. Aryanto, M. Oudkerk, and P. M. van Ooijen, "Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy," *European Radiology*, vol. 25, no. 12, pp. 3685–3695, 2015.

[49] R. Noumeir, A. Lemay, and J. M. Lina, "Pseudonymization of radiology data for research purposes," *Journal of Digital Imaging*, vol. 20, no. 3, pp. 284–295, 2007.

[50] W. Newhauser, T. Jones, S. Swerdloff, W. Newhauser, M. Cilia, R. Carver, A. Halloran, and R. Zhang, "Anonymization of dicom electronic medical records for radiation therapy," *Computers in Biology and Medicine*, vol. 53, pp. 134 – 140, 2014.

[51] T. K. Kim, "Understanding one-way anova using conceptual figures," *Korean journal of anesthesiology*, vol. 70, no. 1, p. 22, 2017.

[52] H. Levene, "Robust tests for equality of variances," *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, pp. 279–292, 1961.

[53] A. Hilton and R. A. Armstrong, "Statnote 6: post-hoc anova tests," *Microbiologist*, vol. 2006, pp. 34–36, 2006.

Appendix A

Data anonymization

Table A.1: DICOM header tags to be treated.

Tag	Name	Action
(0008,0096)	Referring Physician Identification Sequence	Deleted
(0008,1048)	Physician(s) of Record	Deleted
(0008,1049)	Physician(s) of Record Identification Sequence	Deleted
(0008,1050)	Performing Physician's Name	Deleted
(0008,1052)	Performing Physician Identification Sequence	Deleted
(0008,1060)	Name of Physician(s) Reading Study	Deleted
(0008,1062)	Physician(s) Reading Study Identification Sequence	Deleted
(0010,0050)	Patient's Insurance Plan Code Sequence	Deleted
(0010,0101)	Patient's Primary Language Code Sequence	Deleted
(0010,1090)	Medical Record Locator	Deleted
(0010,2180)	Occupation	Deleted
(0010,1002)	Other Patient IDs Sequence	Deleted
(0010,1040)	Patient's Address	Deleted
(0010,1060)	Patient's Mother's Birth Name	Deleted
(0010,0021)	Issuer of Patient ID	Deleted
(0010,2160)	Ethnic Group	Deleted
(0010,21B0)	Additional Patient History	Deleted
(0010,1005)	Patient's Birth Name	Deleted
(0010,2150)	Country of Residence	Deleted
(0010,2152)	Region of Residence	Deleted
(0010,2154)	Patient's Telephone Numbers	Deleted
(0038,0300)	Current Patient Location	Deleted
(0038,0400)	Patient's Institution Residence	Deleted
(0008,0021)	Series Date	Deleted
(0008,002A)	Acquisition DateTime	Deleted
(0008,0031)	Series Time	Deleted
(0008,0032)	Acquisition Time	Deleted
(0008,0081)	Institution Address	Deleted
(0008,0092)	Referring Physician's Address	Deleted

Continued on next page

Table A.1 – DICOM header tags to be treated (continuation).

Tag	Name	Action
(0008,0094)	Referring Physician's Telephone Numbers	Deleted
(0008,1040)	Institutional Department Name	Deleted
(0008,1070)	Operators' Name	Deleted
(0040,A120)	DateTime	Deleted
(0040,A121)	Date	Deleted
(0040,A122)	Time	Deleted
(0008,1010)	Station Name	Deleted
(0008,1030)	Study Description	Deleted
(0008,2111)	Derivation Description	Deleted
(0010,4000)	Patient Comments	Deleted
(0020,4000)	Image Comments	Deleted
(0040,0275)	Request Attributes Sequence	Deleted
(0040,A730)	Content Sequence	Deleted
(0010,0030)	Patient's Birth Date	Deleted
(0010,0010)	Patient's Name	Deleted
(0020,0052)	Frame of Reference UID	Deleted
(0020,0200)	Synchronization Frame of Reference UID	Deleted
(0008,0020)	Study Date	Deleted
(0008,0023)	Content Date	Deleted
(0008,0030)	Study Time	Deleted
(0008,0033)	Content Time	Deleted
(0020,000D)	Study Instance UID	Deleted
(0020,0010)	Study ID	Deleted
(0040,A123)	Person Name	Deleted
(0008,0014)	Instance Creator UID	Deleted
(0008,1155)	Referenced SOP Instance UID	Deleted
(0010,0032)	Patient's Birth Time	Deleted
(0010,1000)	Other Patient IDs	Deleted
(0010,1001)	Other Patient Names	Deleted
(0010,1020)	Patient's Size	Deleted
(0010,1030)	Patient's Weight	Deleted
(0018,1000)	Device Serial Number	Deleted
(0040,A124)	UID	Deleted
(0088,0140)	Storage Media File-set UID	Deleted
(3006,0024)	Referenced Frame of Reference UID	Deleted
(0010,0020)	Patient ID	Pseudo-anonymized
(0008,0018)	SOP Instance UID	Kept
(0008,001A)	Related General SOP Class UID	Kept
(0008,0022)	Acquisition Date2	Kept
(0008,0060)	Modality	Kept
(0008,0070)	Manufacturer	Kept
(0008,1080)	Admitting Diagnoses Description	Kept
(0008,1090)	Manufacturer's Model Name	Kept
(0010,0020)	Patient ID	Pseudo-anonymized

Continued on next page

Table A.1 – DICOM header tags to be treated (continuation).

Tag	Name	Action
(0010,0040)	Patient's Sex	Kept
(0010,1010)	Patient's Age	Kept
(0014,40A2)	Image Quality Indicator Size	Kept
(0018,0024)	Sequence Name	Kept
(0018,0050)	Slice Thickness	Kept
(0018,0080)	Repetition Time	Kept
(0018,0081)	Echo Time	Kept
(0018,0082)	Inversion Time	Kept
(0018,0087)	Magnetic Field Strength	Kept
(0018,0091)	Echo Train Length	Kept
(0018,0093)	Percent Sampling	Kept
(0018,1020)	Software Version(s)	Kept
(0018,1310)	Acquisition Matrix	Kept
(0018,1314)	Flip Angle	Kept
(0018,9005)	Pulse Sequence Name	Kept
(0018,9075)	Diffusion Directionality	Kept
(0018,9076)	Diffusion Gradient Direction Sequence	Kept
(0018,9087)	Diffusion b-value	Kept
(0018,9089)	Diffusion Gradient Orientation	Kept
(0018,9117)	MR Diffusion Sequence	Kept
(0018,9125)	MR FOV/Geometry Sequence	Kept
(0018,9147)	Diffusion Anisotropy Type	Kept
(0018,9240)	RF Echo Train Length	Kept
(0018,9241)	Gradient Echo Train Length	Kept
(0018,9423)	Acquisition Protocol Name	Kept
(0019,000A)	NumberOfImagesInMosaic	Kept
(0019,000B)	SliceMeasurementDuration	Kept
(0019,000C)	B_value	Kept
(0019,000D)	DiffusionDirectionality	Kept
(0019,000E)	DiffusionGradientDirection	Kept
(0019,000F)	GradientMode	Kept
(0019,0027)	B_matrix	Kept
(0019,10bb)	DTI diffusion directions	Kept
(0019,10bc)	DTI diffusion directions	Kept
(0019,10bd)	DTI diffusion directions	Kept
(0019,10d9)	Concatenated SAT	Kept
(0019,10df)	DTI diffusion directions	Kept
(0019,10e0)	DTI diffusion directions	Kept
(0019,0028)	BandwidthPerPixelPhaseEncode	Kept
(0020,000D)	Study Instance UID	Kept
(0020,000E)	Series Instance UID	Kept
(0021,105A)	Diffusion direction	Kept
(0029,1001)	Private Sequence	Kept
(0029,1090)	Private Byte Data	Kept

Continued on next page

Table A.1 – DICOM header tags to be treated (continuation).

Tag	Name	Action
(0054,0081)	Number of Slices	Kept
(2001,1003)	B_value	Kept

Appendix B

GUI screens

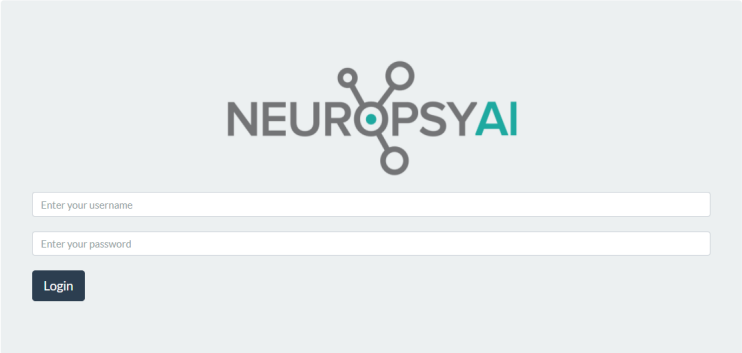


Figure B.1: GUI's login page

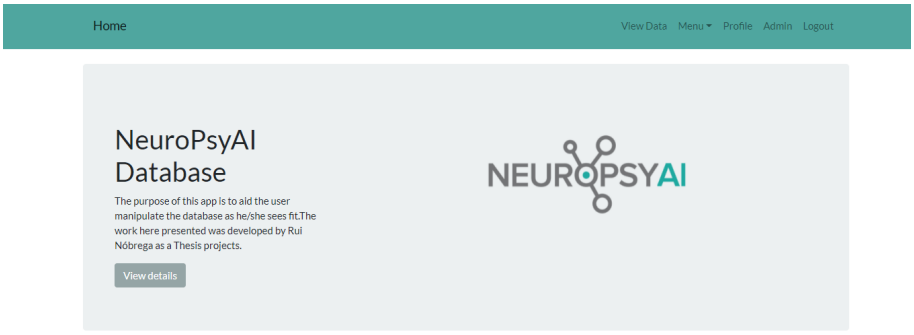


Figure B.2: GUI's home screen

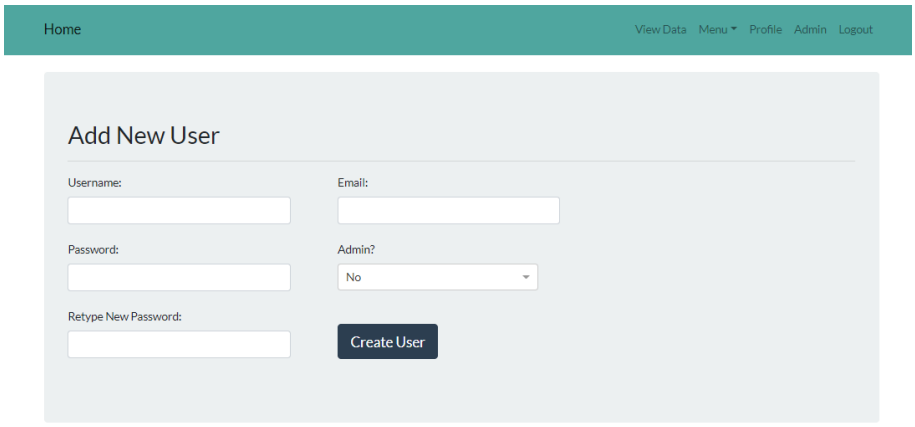


Figure B.3: GUI's add new patient page

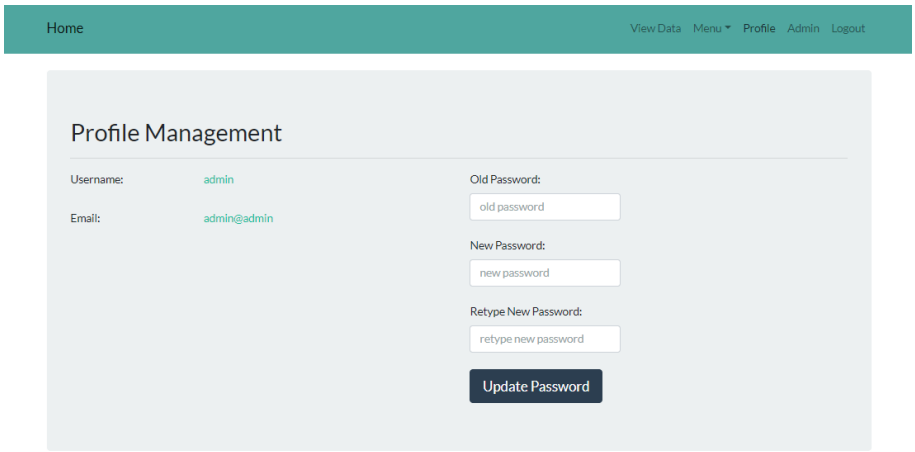


Figure B.4: GUI's update profile page

Appendix C

Database

C.1 Relevant queries

1. What is the distribution of patients in terms of age, gender, hospital and diagnose?

   Query:

```
SELECT
    "Patients_Info"."Patient_ID",
    "Patients_Info"."Age",
    "Patients_Info"."Gender",
    "Patients_Info"."Last_Diagnosis",
    "Patients_Info"."Hospital"
FROM
    public."Patients_Info";
```

2. What is the gender distribution for each known diagnose?

   Query:

```
SELECT
    "Last_Diagnosis",
    "Gender"
FROM
    public."Patients_Info";
```

3. What protocols were used for imaging and how do they represent the population?

   Query:

```
SELECT
    "Value" as Protocol_Name
FROM
    public."Exams_Info"
WHERE
    "Tag" = '(0018,1030)';
```

4. What is the percentage of patients from each hospital/clinic?

   Query:

```
SELECT
    "Value" as Image_Source
FROM
    public."Exams_Info"
WHERE
    "Tag" = '(0008,0080)';
```

5. What is the total number of patients with each diagnose for the different age groups?.

   Query:

```
SELECT
    "Last_Diagnosis",
    "Age"
    Count("Age") as Number_of_patients
FROM
    public."Patients_Info"
GROUP BY
    ("Last_Diagnosis", "Age");
```

6. What does table X contain?

   Query:

```
SELECT
    *
FROM
    table X;
```

Appendix D

External Validation

D.1 Training and testing population

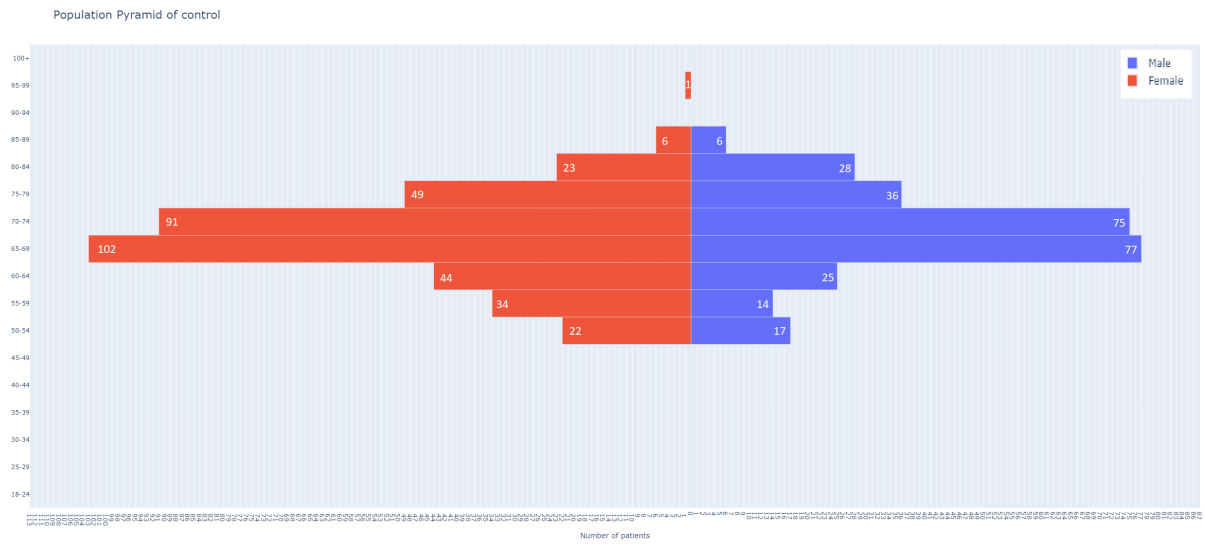


Figure D.1: Distribution of class Control by age group

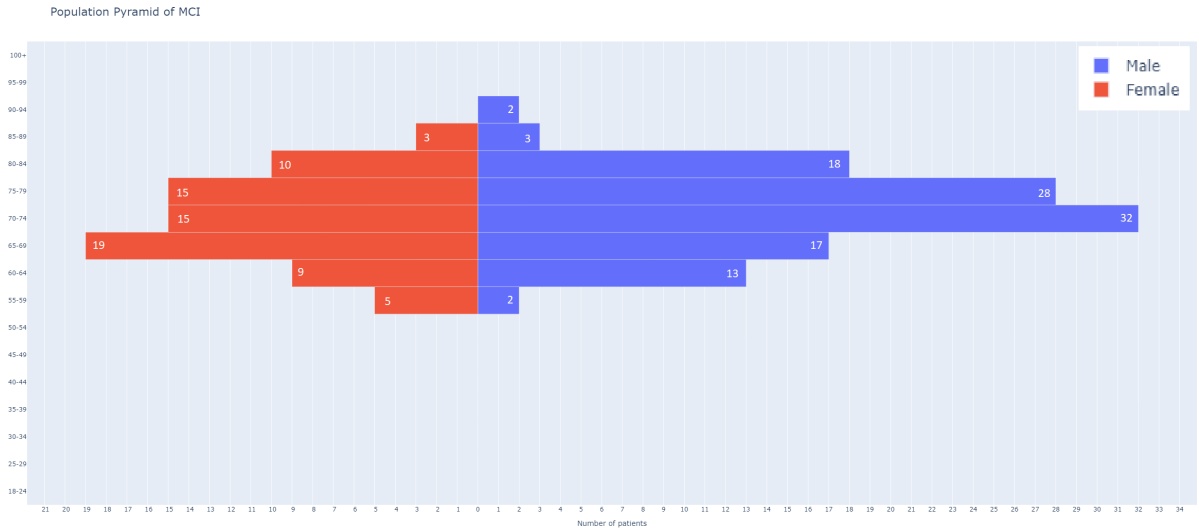


Figure D.2: Distribution of class MCI by age group

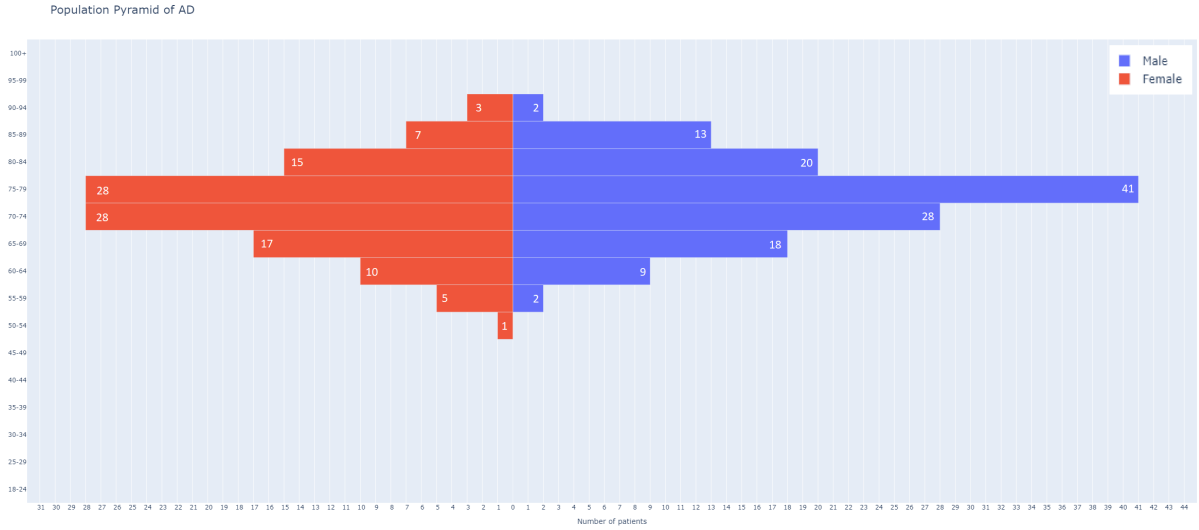


Figure D.3: Distribution of class AD by age group

D.2 Target population

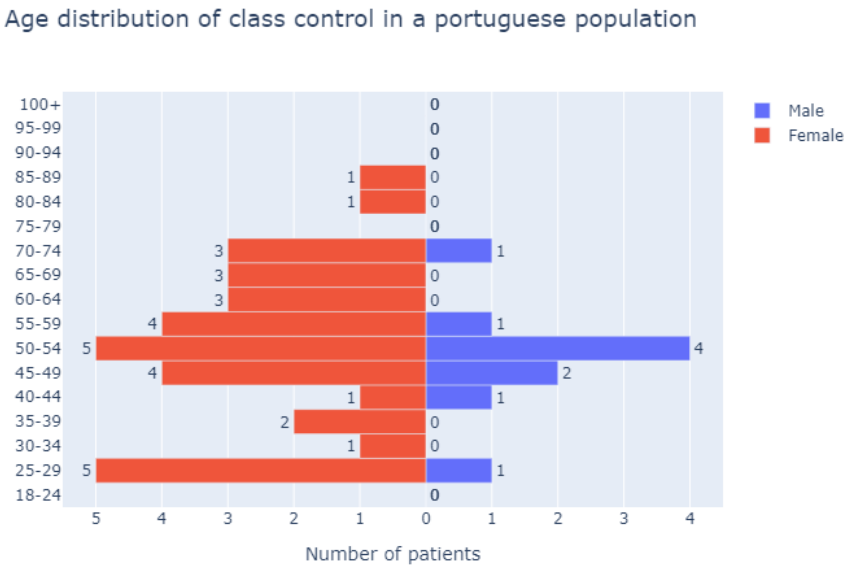


Figure D.4: Distribution of class Control by age group

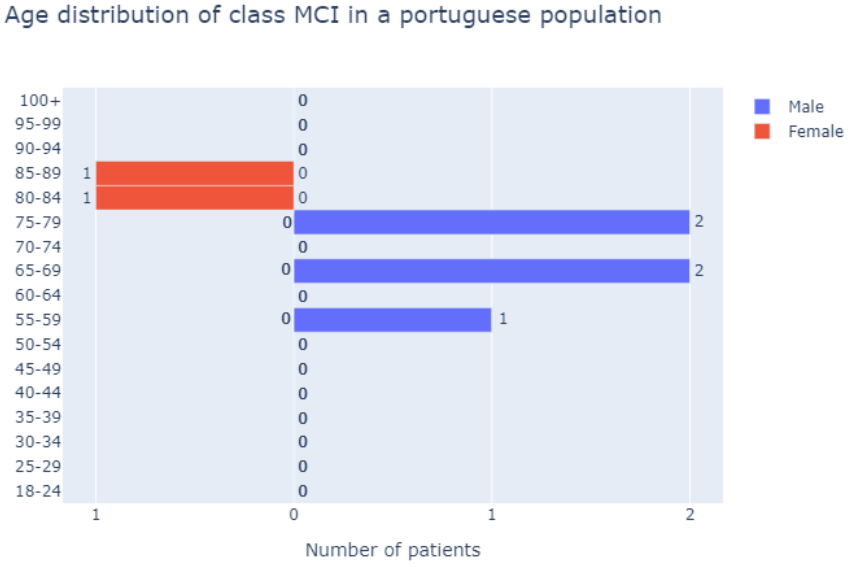


Figure D.5: Distribution of class MCI by age group

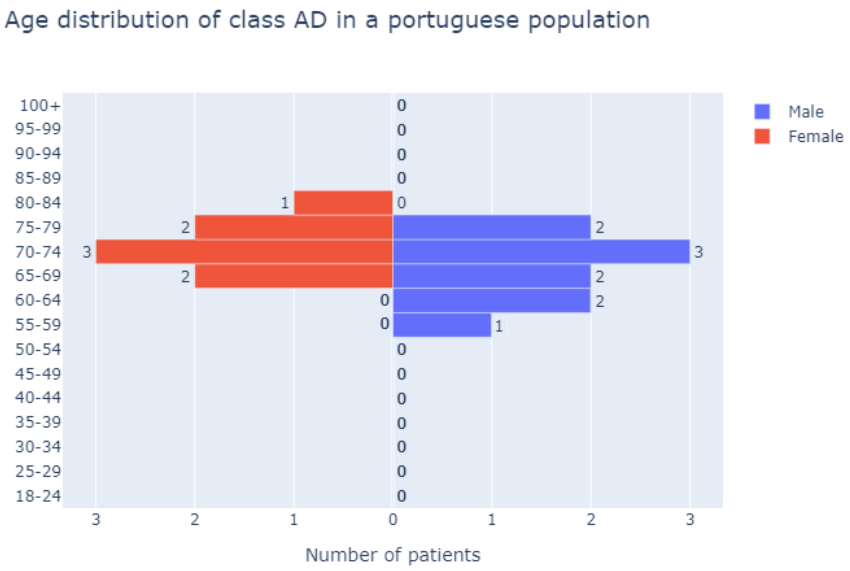


Figure D.6: Distribution of class AD by age group