Predicting the conversion from mild cognitive impairment to Alzheimer’s disease using disease staging

A case study in ADNI data

Ricardo Duarte Ferreira Filipe

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors:
Professor Sara Alexandra Cordeiro Madeira
Doctor Alexandre Valério de Mendonça

Examination Committee
Chairperson: Professor João Miguel Raposo Sanches
Supervisor: Doctor Alexandre Valério de Mendonça
Members of the Committee: Professor Alexandra Sofia Martins de Carvalho

May 2015
I have not failed. I've just found 10,000 ways that won't work.

Thomas A. Edison
Acknowledgments

I would like to express my thanks to Professor Sara Madeira for all her availability to provide guidance and support through all my doubts.

To the kdBio group, I express my many thanks for all the help, availability in clearing my doubts and reviewing my work, and specially, the great team spirit and working environment. This helped me to overcome a lot of the challenges that I faced.

To my dearest colleagues David Machado, Hugo Paixão, João Ramalhinho and Tiago Chaves, I would like to say how grateful I am for sharing this long journey with you, and may another journey follow.

To my long date friends from primary and high school, I express my many thanks for being there in the right moments and for all the support and inspiration.

To my family, I love you and could not complete this journey without you. I would like to dedicate this thesis to my dear aunt Leonor Filipe, who I know is looking after me right now.
Abstract

There is an increasingly necessity to understand the underlying sequence of biological events occurring within the Alzheimer’s disease (AD) pathological cascade. This urge emerges from the fact that AD remains incurable, while facing an exponential increase of prevalence. The known risk factors are being unveiled, but the exact mechanisms triggered by this factors are still unknown. The discovery of the sequence of events preceding the devastating neuropsychological symptoms caused by AD, along with the prediction of subjects who are going to progress to dementia, are of major importance in order to efficiently carry out clinical trials and future research.

In this thesis, we developed a single methodology to merge the estimated patient's disease stage in a data mining procedure to predict conversion to dementia within 3 years, from patients with mild cognitive impairment in the Alzheimer’s Disease Neuroimaging Initiative database. By modelling the longitudinal biomarkers progression, we are able to estimate the disease stage of each patient and extract the long-term growth curves that characterize the AD pathological cascade.

We propose to use the estimated disease stage as a feature for the prediction of conversion, probing multiple classifiers and preprocessing techniques. The augmented feature set led to an overall increased power of generalization across the different applied classifiers. We observed considerable improvements in the validation results when incorporating the disease stage, rather than learning just the baseline dataset.

Keywords

Alzheimer’s Disease, MCI, Data Mining, Prognosis, Longitudinal Model, GRACE
Resumo

A crescente necessidade de inferir a sequência de eventos biológicos que caracteriza a cascata patológica da doença de Alzheimer (DA) surge perante um aumento exponencial na prevalência desta doença e do facto de continuar incurável. Os factores de risco estão a ser desvendados, mas os mecanismos desencadeados por estes factores ainda permanecem desconhecidos. Desvendar a sequência de eventos que precede os devastadores sintomas neuropsicológicos causados pela DA e prever o prognóstico de demência são dois importantes objectivos para promover a eficácia dos testes clínicos e futura investigação.

Nesta dissertação, desenvolveu-se uma metodologia para incluir uma estimativa da fase da doença de cada sujeito num modelo para previsão da conversão para demência dentro de 3 anos em pacientes com deficiência cognitiva ligeira, usando dados da Alzheimer's Disease Neuroimaging Initiative. Ao modelar a progressão longitudinal dos biomarcadores, nós conseguimos estimar as fases da doença dos pacientes e extrair curvas de crescimento a longo-prazo que caracterizam a cascata patológica da DA.

Com esta dissertação, propomos a incorporação da fase da doença, como atributo para a previsão da conversão, testando múltiplos classificadores e técnicas de pré-processamento. O conjunto de atributos proposto contribuiu para um aumento global do poder de generalização nos diferentes classificadores usados. Observamos que os resultados da validação usando o atributo proposto são consideravelmente melhores do que sem esse atributo.

Palavras Chave

Doença de Alzheimer, Défice Cognitivo Ligeiro, Classificação, Prognóstico, Modelos Longitudinais, GRACE
Contents

1 Introduction 1
  1.1 Context and Motivation ................................................. 2
  1.2 Goals ................................................................. 3
  1.3 Contributions .......................................................... 4
  1.4 Thesis Outline .......................................................... 4

2 Background 7
  2.1 Alzheimer's Disease ..................................................... 8
    2.1.1 AD Pathological Cascade ......................................... 9
  2.2 Biomarkers of AD ....................................................... 12
    2.2.1 Proteomic Biomarkers ............................................. 12
    2.2.2 Neuroimaging Biomarkers ........................................ 13
    2.2.3 Neuropsychological Tests ....................................... 15
  2.3 AD Neuroimaging Initiative ........................................... 17
  2.4 Related Work ........................................................... 21
    2.4.1 Longitudinal Models ............................................. 21
      2.4.1.A Long-term Progression Curves with Short-term Data ........ 22
      2.4.1.B Alzheimer's Disease Event Ordering ......................... 23
    2.4.2 Data Mining with ADNI Data .................................... 25

3 Methods 31
  3.1 Proposed Approach .................................................... 32
  3.2 Data Preprocessing .................................................... 33
    3.2.1 Dataset Statistical Description ................................ 34
    3.2.2 Instance Labelling Variable .................................... 36
    3.2.3 Extracting Progression to AD .................................... 39
  3.3 Computing Learning Examples ........................................ 42
  3.4 GRACE Models ......................................................... 44
    3.4.1 GRACE Preprocessing ............................................. 47
    3.4.2 Computing GRACE+ Learning Examples ........................... 49
  3.5 Classification Methodology ........................................... 50
    3.5.1 Classification Preprocessing Techniques ........................ 52
# List of Figures

2.1 Proposed mechanisms underlying AD ........................................... 10  
2.2 Sequential cleavage of APP ....................................................... 11  
2.3 Timeline of the 3 phases of the ADNI. ................................. 18  
2.4 Model of biomarker deterioration dynamics. ......................... 22  

3.1 Proposed methodology pipeline ........................................ 33  
3.2 Preprocessing routine implemented to extract, clean and correct data 34  
3.3 Random sample of the original dataset. .......................... 36  
3.4 Example of an anomalous assigned visit code. ..................... 37  
3.5 Function of assigned visit code given an exam date for an exemplifying set of pivots dates. 38  
3.6 Example of 3 different visit code corrections. .................... 39  
3.7 Histogram of visit codes after correction. ......................... 40  
3.8 Correction of DXChange based on the comparison between the dataset exam date and DXSum exam date. ................................. 41  
3.9 Correction of DXChange based on the diagnosis at baseline extracted from the file arm. 41  
3.10 Correction of DXChange based on the assumption of a continuous deterioration of diagnosis status. ............................ 42  
3.11 Diagram representing the number of data entries restored and removed when applying the proposed corrections. 42  
3.12 Class labelling scheme. ...................................................... 43  
3.13 Plot of the number of instances versus the number of years for the temporal window. 44  
3.14 Scheme of the first two GRACE iterations. ...................... 46  
3.15 Histogram of the different categories used in the construction of the GRACE models. 47  
3.16 Data sample in the required format to enter GRACE algorithm. 48  
3.17 Percentile Transform Function ........................................... 49  
3.18 Processing of mapping γ and tγ values for points in the test fold. 50  
3.19 Classification methodology data flow ........................................ 51  
3.20 Missing values percentage for each variable. ..................... 54  
3.21 Class distributions after the used SMOTE percentages, across all the folds and seeds. 55  

4.1 Progression Curves obtained from the GRACE model using the training dataset .......... 59  
4.2 Baseline Results across different classifiers with different SMOTE percentages. 61
4.3 The 10-fold CV results using Baseline and GRACE+ datasets, across all the classifiers.

4.4 The 10-fold CV results using Baseline and GRACE+ datasets, across all the classifiers and including all the various preprocessing combinations.

4.5 Validation results using Baseline and GRACE+ datasets for the best trained classifier, Naïve Bayes without preprocessing.

4.6 Validation results using BL and GRACE+ datasets, across the top classifiers which yielded the highest AUCW in training.

4.7 Validation results using Baseline and GRACE+ datasets, across all the classifiers, including all preprocessing techniques combination for each classifier. Some of the poorer results are not shown in order to adjust the scale of visualization.

4.8 Results across multiple classifiers with different SMOTE percentages and the other techniques as the optimized in training.

4.9 Results across multiple classifiers with and without missing values imputation (mvi) and the other techniques as the optimized in training.

4.10 Feature selection frequency applying the two employed FS techniques.

4.11 Results across multiple classifiers without feature selection (fs), with correlation-based fs and with NB wrapper and the other techniques as the optimized in training.

B.1 Smooth densities estimates for the set of numeric continuous features and Histograms of discrete features.

B.2 Percentage of times each feature was selected from 4 feature different selection methods.

C.1 Convergence analysis of the model constructed from the whole training set. The process is stopped by reaching the tolerance level ($10^{-3}$) in the 9th iteration. RSS=Residual Sum of Squares.

C.2 GRACE progression curves and shifted dataset in iterations 1 and 2, using the training dataset (75% of the whole dataset).

C.3 GRACE progression curves and shifted dataset in iteration 5, using the training dataset (75% of the whole dataset).

C.4 The $\gamma$ estimations in iteration 1, using the training dataset (75% of the whole dataset).

C.5 $\gamma$ estimations in iterations 5 and 9, using the training dataset (75% of the whole dataset).

C.6 GRACE progression curves, from the several CV folds, including the variance obtained from the different datasets.

D.1 Top 7 classifiers with highest AUCW from the whole set of classifier/preprocessing combinations using the BL dataset.

D.2 Top 11 classifiers with highest AUCW from the whole set of classifier/preprocessing combinations using the GRACE+ dataset.

E.1 Scatterplot of the estimated $\gamma_{RAVLT}$ vs $\gamma_{FAQ}$. 
List of Tables

2.1 Number of participants recruited in each phase and types of data. . . . . . . . . . . . . . . . 18
2.2 Inclusion/Exclusion criteria at screening for new participants in ADNI 2. . . . . . . . . . . . 19
2.3 Techniques applied regarding different methodology aspects of the studies presented . . . 26

3.1 Computing features from files variables. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35
3.2 Number of data entries and number of different subjects in each of the diagnosis categories. 36
3.3 Possible Visit Codes labels and respective description. . . . . . . . . . . . . . . . . . . . . 37
3.4 Partial residuals for each target parameter and the respective conditional expectations. . . 45
3.5 Training set characteristics. This dataset corresponds to 75% of the whole dataset and it is
   from this dataset that the CV creates the 10 folds. . . . . . . . . . . . . . . . . . . . . . . . 51
3.6 Test set characteristics. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52
3.7 Parameters Grid Search. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52

4.1 10-fold CV results using Baseline and GRACE+ datasets, across all the classifiers. . . . . . 63
4.2 Differences of Specificity and Sensitivity between GRACE+ and BL and differences be-
   tween Sensitivity and Specificity within each dataset. . . . . . . . . . . . . . . . . . . . . . 64
4.3 10-fold CV results using Baseline and GRACE+ datasets, across all the classifiers and
   including all the various preprocessing combinations. . . . . . . . . . . . . . . . . . . . . . 65
4.4 Differences between the training results and the validation results. . . . . . . . . . . . . . . 67
4.5 Validation results using Baseline and GRACE+ datasets, across all the classifiers, including
   all preprocessing techniques combination for each classifier. . . . . . . . . . . . . . . . . 69

A.1 Number of subjects for each diagnosis category on the training dataset used in each of the
   presented studies. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A-2
A.2 Features incorporated in each of the presented studies. . . . . . . . . . . . . . . . . . . . . A-3

B.1 Characteristics of the Learning Dataset. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . B-2
B.2 Distribution of subjects according to the number of instance per subject. . . . . . . . . . . . B-2
B.3 Top 7 selected features in 10-fold CV Gain Ratio and Info Gain rankings. . . . . . . . . . . B-4

D.1 Optimized parameters for each classifier’s best performance. . . . . . . . . . . . . . . . . . D-3
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>Alzheimer's Association</td>
</tr>
<tr>
<td>AD</td>
<td>Alzheimer's Disease</td>
</tr>
<tr>
<td>ADAS (ADAS-Cog)</td>
<td>Alzheimer's Disease Assessment Scale (cognitive subscale)</td>
</tr>
<tr>
<td>ADCS</td>
<td>Alzheimer's Disease Cooperative Study</td>
</tr>
<tr>
<td>ADGC</td>
<td>Alzheimer's Disease Genetics Consortium</td>
</tr>
<tr>
<td>ADI</td>
<td>Alzheimer's Disease International</td>
</tr>
<tr>
<td>ADNI</td>
<td>Alzheimer's Disease Neuroimaging Initiative</td>
</tr>
<tr>
<td>ANART</td>
<td>American National Adult Reading Test</td>
</tr>
<tr>
<td>ApoE</td>
<td>Apolipoprotein E</td>
</tr>
<tr>
<td>ApoE4</td>
<td>Apolipoprotein E allele 4</td>
</tr>
<tr>
<td>APP</td>
<td>Amyloid Precursor Protein</td>
</tr>
<tr>
<td>BL</td>
<td>Baseline</td>
</tr>
<tr>
<td>CDR</td>
<td>Clinical Dementia Rating</td>
</tr>
<tr>
<td>CDRSB</td>
<td>CDR-Sum of Boxes</td>
</tr>
<tr>
<td>CN</td>
<td>Normal Controls</td>
</tr>
<tr>
<td>CSF</td>
<td>Cerebrospinal fluid</td>
</tr>
<tr>
<td>CV</td>
<td>Cross-Validation</td>
</tr>
<tr>
<td>CVLT</td>
<td>California Verbal Learning test</td>
</tr>
<tr>
<td>DAD</td>
<td>Disability Assessment Scale for Dementia</td>
</tr>
<tr>
<td>DTI</td>
<td>Diffusion Tensor Imaging</td>
</tr>
<tr>
<td>EDC</td>
<td>Electronic Data Capture</td>
</tr>
<tr>
<td>EMCI</td>
<td>Early Mild Cognitive Impairment</td>
</tr>
<tr>
<td>FAQ</td>
<td>Functional Activities Questionnaire</td>
</tr>
<tr>
<td>FDG</td>
<td>2-fluorodeoxy-D-glucose</td>
</tr>
<tr>
<td>fMRI</td>
<td>Functional MRI</td>
</tr>
<tr>
<td>GDS</td>
<td>Geriatric Depression Scale</td>
</tr>
<tr>
<td>GRACE</td>
<td>Growth models by Alternating Conditional Expectation</td>
</tr>
<tr>
<td>LM</td>
<td>Logical Memory</td>
</tr>
<tr>
<td>LMCI</td>
<td>Late Mild Cognitive Impairment</td>
</tr>
<tr>
<td>MCI</td>
<td>Mild Cognitive Impairment</td>
</tr>
<tr>
<td>MMSE</td>
<td>Mini-Mental State Examination</td>
</tr>
<tr>
<td>MoCA</td>
<td>Montreal Cognitive Assessment</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>NFT</td>
<td>Neurofibrillary Tangles</td>
</tr>
<tr>
<td>NIA</td>
<td>National Institute of Aging</td>
</tr>
<tr>
<td>NINCDS-ADRDA</td>
<td>National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association</td>
</tr>
<tr>
<td>NL</td>
<td>Normal Controls</td>
</tr>
<tr>
<td>NM</td>
<td>Neuropsychological Measurements</td>
</tr>
<tr>
<td>PET</td>
<td>Positron emission Tomography</td>
</tr>
<tr>
<td>PIB</td>
<td>Pittsburgh compound B</td>
</tr>
<tr>
<td>p-tau</td>
<td>Phosphorylated tau</td>
</tr>
<tr>
<td>RAVLT</td>
<td>Rey Auditory Verbal Learning Test</td>
</tr>
<tr>
<td>SMC</td>
<td>Significant Memory Concern</td>
</tr>
<tr>
<td>SMOTE</td>
<td>Synthetic Minority Oversampling Technique</td>
</tr>
<tr>
<td>sMRI</td>
<td>Structural MRI</td>
</tr>
<tr>
<td>UPS</td>
<td>Ubiquitin-Proteasome System</td>
</tr>
<tr>
<td>VOSP</td>
<td>Visual Object and Space Perception</td>
</tr>
<tr>
<td>WAIS</td>
<td>Wechsler Adult Intelligence Scale</td>
</tr>
</tbody>
</table>
1 Introduction

Contents

1.1 Context and Motivation .............................................. 2
1.2 Goals ........................................................................... 3
1.3 Contributions ............................................................... 4
1.4 Thesis Outline ............................................................. 4
1.1 Context and Motivation

We humans often forget that all the things we learn, the emotions we feel, the language we produce, basically everything we experience is a mirror of a constantly changing and super complex neuronal circuit. In fact, what we are, our personality, dreams and desires can be seen indeed as the product of our brain, whether being voluntarily remembered or a matter of the subconscious. All the metaphysic hypotheses related to our mind and cognitive functions have, however, an inescapable connection to a real biological system.

Neuropsychological disorders are becoming a cruel reality to a great deal of worldwide populations. Due to poor understanding of the underlying mechanisms and a significant overlap with normal ageing complaints, sociological preconceptions and misunderstandings associated with expectations arise. This is a problem often underrated and trivialized by government healthcare sectors, by physicians and even by family members and friends. The fact is that the prevalence of this type of disorders is growing at an epidemic pace. As we observe increasing life expectancy, decline in physical activities and unbalanced eating habits, it is becoming increasingly urgent to tackle this issue, specifically aiming at one of most devastating neuropsychological disease as well as one of the top causes of death in developed countries [1], Alzheimer's Disease (AD).

AD is a gradually progressive syndrome that mainly affects memory function, ultimately culminating in a dementia state where all cognitive functions are affected. Before reaching the last AD stage, patients undergo an intermediary stage which encompasses the initial growth, termed Mild Cognitive Impairment (MCI). AD is currently the most common cause of dementia, is at the present time incurable and the therapeutic procedures aiming to slow down the progression of disease are still crudely developed.

The current understanding of the disease pathological cascade states that AD starts long before the onset of the disease, being gradually built a number of offenses to the normal brain functioning. Accordingly, the currently well established amyloid cascade hypothesis [2] suggests that AD begins with abnormal β amyloid (Aβ) protein folding which in turn initiates all the other biological events that lead to AD. Jack et.al [3, 4] interpreted the current empirical biomarker evidence and published a hypothetical model of the biomarker dynamics on AD, postulating that biomarkers deterioration from normal to abnormal values occur in an orderly systematic fashion. The rising of cerebrospinal fluid (CSF) Aβ concentration levels to abnormal values, resulting from an unusual accumulation of extracellular senile plaques, is expected to be observed in an initial stage, followed by the detection of irregular Aβ imaging patterns, reduction in tau-protein concentration, due to the abnormal folding into neurofibrillary tangles (NFT), and finally by neurodegeneration along with, at a later point in time, functional/cognitive disability. Neurodegeneration may be measured by assessing brain region volumes, with magnetic resonance imaging (MRI) techniques, and by assessing brain activity, with positron-emission tomography (PET) using specific radioactive tracers. A decline in cognition is followed after a variable time, depending on brain resilience and cognitive reserve, and may be evaluated by the performance on clinical neuropsychological tests.

Nevertheless, there is still contradictory evidence suggesting distinct interpretations on this proposed sequence of events. Finding a coherent and validated model of the disease progression along time, often referred as longitudinal changes, is crucial to find a cure, by improving clinical trials follow-up and by
refining diagnosis and prognosis predictive models.

Facing the numerous efforts approaching AD in the last few years, the AD Neuroimaging Initiative (ADNI\textsuperscript{1}) was designed in order to standardized data collection methods and support data sharing. ADNI is an ongoing project that has been collecting multi-domain data from multiple American centres in the past 10 years. One of their major focus is to tackle the problem of heterogeneous data collection by standardizing all the clinical and imaging procedures, and also to provide worldwide access to data without embargo.

Several studies exploring the ADNI dataset can be grouped into two main types: 1) data mining approaches, in which the main goal of the study is to develop a tool to accurately classify a patient with a diagnosis, or to predict with confidence the prognosis and 2) descriptive approaches, in which the main goal is to conceptualize and construct models of disease progression over time (longitudinal models) supported by empirical evidence. Recently, in the later context, Donohue et al.[5], based on the concept proposed by Jack et al.[3, 4], constructed longitudinal models of biomarker severity progression across a long-term disease timeline. Using data from shorter follow-up periods, he was able to infer the inherit disease stage of each patient and, accordingly, apply a time-shift data transformation to estimate curves of progression through a 2-decades timespan.

### 1.2 Goals

When a prognosis classifier is trained from a dataset with the purpose of optimizing model evaluation metrics, ranking the classifier according to its power of generalization or prediction accuracy, the outcomes are often uninterpretable. The underlying processes of a given classifier are blind to known concepts and models of the disease and therefore, present very important results that are unbiased and may come to invalidate some well-established hypothesis, but often do not help in further understanding the underlying mechanisms of the disease. With this thesis, we aim to develop a data mining methodology that takes into account current model-based knowledge to investigate prediction of conversion from MCI to AD. Namely, the goal of this thesis is to assess the predictive power of features depicting the patient’s disease stage using multi-domain ADNI data in a proposed data mining approach to predict progression from MCI to AD, within 3 years. The disease stage is extracted by fitting biomarker data to multiple progression curves covering a wide disease timespan, expanded from short follow-up periods of numerous patients. The algorithm used for modelling long-term progression curves was constructed from the work published by Donohue et al.[5]. With this algorithm and implemented data mining techniques, we aim at three main objectives:

1. Modelling biomarker dynamics in a long-term disease timeline, in order to describe the pathophysiological cascade underlying AD, by providing a picture on the way biomarkers severity develops through the course of the disease;
2. Training a prognosis classifier to predict conversion from MCI to AD using key-biomarkers from ADNI data, as well as model-extracted features of disease stage. This includes:

\footnote{\textbf{ADNI website:} \url{http://adni.loni.usc.edu/}}
(a) Identifying the specific classifier and the combination of preprocessing techniques, including missing values imputation, feature selection and data balancing, that yield the best performance results for the given dataset;
(b) Evaluating the robustness and generalization power of such classifier/preprocessing approach, in a validation dataset.

3. Assess the predictive power and overall benefits of model-extracted features of disease stage when introduced into the classifiers.

1.3 Contributions

The contributions of this work can be summarized in the following:

1. A wide literature review covering recent studies using the ADNI dataset regarding the development/application of descriptive and data mining approaches;
2. Development of a preprocessing routine to detect and correct dataset inconsistencies;
3. Development of a data mining methodology for classifying patients according to the prognosis of conversion from MCI to AD within 3 years, considering multiple classifiers and preprocessing techniques;
4. A detailed study on the separability of AD converters from non-converters using disease stage, and on the effects of incorporating disease stage features in the classifiers results.
5. Analysis of the influence of different preprocessing techniques in the classifiers results;

1.4 Thesis Outline

This dissertation is outlined as follows:

Chapter 2 - Background: In this chapter, we will present the necessary concepts that one should grasp in order to perceive the current context of AD research, the well-established concepts and the emerging hypothesis and thus fully understand the basis upon which the work developed in this thesis was constructed. With this purpose, we group this chapter in the following four sections:

Section 2.1 - Alzheimer’s Disease: we present AD main characteristic symptoms and figures, such as the prevalence and worldwide costs. The research efforts and the current therapeutics are also discussed. The current understanding of the biological mechanisms underlying AD is presented in Section 2.1.1.

Section 2.2 - Biomarkers of AD: we present the main biomarkers that are used to detect AD, i.e. Proteomic, Imaging and Neuropsychological measurements, and analyse their usefulness in modern AD research.

Section 2.3 - AD Neuroimaging Initiative (ADNI): we explain in detail the project from which we are extracting the dataset, its main objective, the data being acquired and data extraction protocols.

Section 2.4 - Related Work: we present the outcome of a wide literature review on the different recent studies performed using the ADNI dataset. We group the studies according to their main
purposes: descriptive disease modelling or AD diagnosis/prognosis prediction.

Chapter 3 - Methods: In this chapter, we will explain in detail the methodology employed in this dissertation, which includes:

Section 3.1 - Proposed Approach: the big picture of the entire methodology is illustrated.
Section 3.2 - Data Preprocessing: we present the dataset characteristics and discuss the implemented data preprocessing along with the reasons for their implementation.
Section 3.3 - Computing Learning Examples: we describe the process underlying the creation of data instances from separated data files and present the characteristics of the classifiable dataset.
Section 3.4 - GRACE Models: the GRACE algorithm, the respective preprocessing and the process of computing learning examples with GRACE features are explained in detail.
Section 3.5 - Proposed Classification Methodology: the classification methodology is explained in detail, including the dataset workflow, in training and validation, the employed preprocessing techniques and the model evaluation metrics used to rank classifiers.

Chapter 4 - Results and Discussion: The presented results are grouped in the following outline:

Section 4.1 - GRACE curves: we show the progression curves resulting from the constructed GRACE models and also the classes separability using the GRACE parameters.
Section 4.2 - Baseline vs GRACE+ Results: we present and compare the results between using the augmented dataset against the baseline dataset.
Section 4.3 - Preprocessing Results: we present the results of a detailed study on the effects of the employed preprocessing techniques in the final results.
Section 4.4 - Summary: we summarize the different aspects of the results presented in this chapter.

Chapter 5 - Conclusions and Future Work: In this chapter, we draw the final conclusions and point to potential future work that can be performed from this thesis.

Appendix A - Related Work (Complementary Information): we present two tables characterizing the presented related work accordingly to the used dataset and feature set.
Appendix B - Dataset Description and Statistical Analysis: We present statistical descriptive analysis of the classifier dataset and a exploratory analysis on the features’ informative value.
Appendix C - GRACE Outcomes: The results from the GRACE iterations are shown, along with the 10-fold CV resulting GRACE curves.
Appendix D - Classification (Complementary Results): We present some results complementary to the analysis performed in this work.
Appendix E - Future Work: Here, we explore in more detail the proposed future work ideas.
2

Background

Contents

2.1 Alzheimer’s Disease ........................................ 8
2.2 Biomarkers of AD ........................................ 12
2.3 AD Neuroimaging Initiative .............................. 17
2.4 Related Work .............................................. 21
2.1 Alzheimer’s Disease

Alzheimer’s disease (AD) is a chronic neurodegenerative disease, characterized by a slow progressive deterioration of memory function, from an early stage of subtle cognitive symptoms to late-stage dementia, being the most common cause of dementia, accounting for an estimated 60% to 80% of the total number of cases [6]. This disease is caused by loss of neurons and synapses in the cerebral cortex and certain subcortical regions, resulting in gross atrophy of the affected regions. Through the course of disease progression, the symptoms may vary and can include besides memory, other cognitive functions decay, such as language, orientation, mood or motivation. Moreover, the progression of AD is now understood as a cascade of pathological events that precedes the onset of the disease by more than 2 decades, thus leading to a wide range of disease states with different characteristics, ordered differently from patient to patient. This heterogeneity leads to a symptomatic profile with a very large spectrum of possible outcomes, such as short-term memory loss, impairments in semantic memory, apathetic behaviour, difficulties with language and perception or problems with executive functions, such as planning, attentiveness and abstract thinking. Nevertheless, it has been detected a pattern distinct from normal ageing and dementia, in intermediary stages of patients who transit to AD, termed Mild Cognitive Impairment (MCI). Subjects with MCI have a memory impairment beyond what is expected for age and education, yet are not demented. Since its introduction [7], MCI and its possible connection to AD have been intensively studied from many perspectives including clinical, imaging, genetic, pathological and epidemiological.

There are currently approximately 33.9 million individuals with AD worldwide, and, due to demographic changes and the increase of average life expectancy, is expected an epidemic exponential growth with prevalence triplicating over the next 40 years [8]. According to ADI’s World Alzheimer Report [9, 10], the total estimated worldwide costs of dementia in 2010 was US$640 billion, based on demographics, the ADI report foresees an 85% increase in cost by 2030, with developing countries bearing an increasing share of the economic burden. These facts and figures emphasize the importance of preventing this debilitating disease to reduce the burden on society as well as easing suffering of those affected and their families. In fact, it can already be observed worldwide efforts to fight the increasingly inflation of AD prevalence, for instance, the United States National Alzheimer’s Project Act signed into law in 2011, which required a national plan to accelerate innovative treatment findings and research in prevention of AD.

Regarding support for AD research, several entities can be named, such as 1) the Alzheimer’s Association, the world’s leading voluntary health organization in AD support and research; 2) the US National Institute of Aging (NIA) projecting, cooperating and providing grants to several projects, including the Alzheimer’s Disease Neuroimaging Initiative (ADNI), the Alzheimer’s Disease Cooperative Study (ADCS), focused on the development of compounds to improve AD patients quality of life, and the Alzheimer’s Disease Genetics Consortium (ADGC), directed to genome-wide association studies to identify genes associated with an increased AD risk; or 3) the Alzheimer Europe, an organization of 36 Alzheimer associations from 31 countries across Europe, supporting projects like Pharmacog, focusing on increasing the
ability to predict new pharmacological treatments from laboratory studies and clinical models. Other AD research drivers involve challenges like the Alzheimer’s Disease Big Data DREAM Challenge, supported by Sage Bionetworks, on which participants were propelled to rapidly and accurately identify biological features to predict AD.

Although research in the last 30 years has gained momentum and revealed very important aspects of the disease, nowadays, there is no treatment to cure or stop AD progression. Dozens of drugs and pharmacological therapies aiming at slowing or stopping neuronal malfunction and degeneration are being studied around the world. These studies include a recent systematic review of nine randomized clinical trials with a total of 5149 persons with MCI, which reported no effect of cholinesterase inhibitors (donepezil, galantamine and rivastigmine) on cognition or slowing the progression to AD within 3 years, among other studies, such as a similar systematic review which suggested an unsuccessful role of Vitamin E in benefiting the treatment of MCI. For severe AD patients, therapeutic trials with cholinesterase inhibitors and memantine have been conducted and consequently, treatment with memantine has been approved by numerous drug agencies and donepezil has been approved by the US FDA [11]. Furthermore, non-pharmacological therapies are being proven to possibly benefit MCI patients, including cognitive training and physical exercise [7].

The current strategies are focusing on slowing down the disease progression and sustaining the patients’ level of cognitive and functional ability to ensure a tolerable quality of life for them and their respective caregivers. Therapy research is also shifting to early stages of the disease towards preclinical AD in order to increase efficiency on some treatments and to harvest the benefits of prevention [12]. However, with the estimation of a worldwide dementia increase and with the insufficient and ineffective current ways of facing AD, further investment and stimulation in research is urgently necessary to slow down this devastating disorder.

2.1.1 AD Pathological Cascade

Alois Alzheimer, in 1907, observed the abnormal agglomeration of unknown metabolites in the neocortex and hippocampus of the autopsied brain of a middle-aged woman affected for years with uncommon deficits in memory and cognitive functions. However, over 80 years elapsed before sufficient evidence had been accumulated to identify the principle components that formed those unusual aggregations. A milestone in neurodegenerative disorder research was indeed the identification of those metabolites: the AD-characteristic senile plaques and neurofibrillary tangles.

Just over 20 years ago, the infectious protein particle in scrapie sheeps, termed prion, was identified as the transmissible agent of that pathology. The discovery of the prion’s property to promote the aggregation of plaque-like amyloid deposits, coupled with the detection of the neurofibrillary building blocks of neurodegenerative disorders, led to the identification of the patterns that would open the path for the modern molecular era of research on neurological disorders.

Most neurodegenerative diseases have a significant overlap in the pathophysiological cascade [13]. Currently, a common mechanism of neurodegenerative diseases is linked to two critical events: 1) a pathological aggregation of misfolded proteins and 2) an accumulation of fibrillar amyloid deposits in certain vulnerable regions of the central nervous system. Mediated by genetic and environmental factors,
highly soluble proteins are gradually converted into insoluble filamentous polymers that accumulate in a disease- and protein-specific manner as fibrillar amyloid deposits.

AD-specific pathology is characterized by 1) the agglomeration of \( \beta \) amyloid (A\( \beta \)) senile plaques, resulting from a deficient cleavage of the amyloid precursor protein (APP), and 2) neurofibrillary tangles (NFTs), formed from aberrantly phosphorylated tau, a microtubule-associated protein (Fig. 2.1) [14]. In fact, it is becoming increasingly evident that these misfolded and aggregated protein complexes are indeed directly related to disease pathogenesis instead of simply being intermediary steps of AD neuropathological path. This is the basis of the amyloid hypothesis [2], which postulates that a serial causal sequence of events results in dementia due to AD and is initiated with abnormal elevations in A\( \beta \), causing a posterior abnormal tau hyperphosphorylation.

**Figure 2.1:** Proposed mechanisms underlying AD, including the pathological cascade of events from the protein misfolding and fibrillization to the deposition of aggregated tau filaments and senile plaques (fibrillar A\( \beta \)). The amyloid hypothesis suggests that \( \beta \) fibrils lead to deposition of extracellular senile plaques and also intracytoplasmic NFTs. On an early stage of misfolding, molecular chaperones, Proteasomes and Phagosome/Lysosome systems act to invert the pathological unbalance and the sequence of biological events that lead to plaque deposition. [15]

Regarding A\( \beta \) senile plaques, when investigating the mechanisms underpinning its abnormal accumulation, the identification of mutations in the genes that encode APP and presenilins 1 and 2 led to the discovery of a key aspect for nonamyloidogenic to shift to amyloidogenic A\( \beta \) species: the type of proteolytic enzymes involved in the cleavage of APP (Fig. 2.2). It was also found that most presenilin mutations alter the specific APP C-terminal cleavage site of \( \gamma \)-secretase, thus promoting the generation of A\( \beta_{42} \).

\( \beta \) amyloid auto-oxidation releases free reactive oxygen and reactive nitrogen species causing neuronal damage [16], which in turn will lead to massive neuronal tissue degeneration and brain atrophy, if not decelerated. However, cells have already adapted a sophisticated control mechanism to slow down the accumulation of misfolded and aggregated proteins, whether by promoting proper folding with molecular chaperones or by degrading remains of misfolded aggregates, with the ubiquitin-proteasome system (UPS) and the phagosome-lysosome system (Fig. 2.1) [14]. Ubiquitin mutations or defects may impair removal of the insoluble neurofibrillary material that aggregates into tangles and senile plaques. Interestingly, A\( \beta \) is one of the factors that could disable proteasome activity.
Figure 2.2: The sequential cleavage of APP by $\beta$-secretase (BACE-1) followed by $\gamma$-secretase, contrary to the cleavage in the middle of the APP mediated by $\alpha$-secretases, produces potentially amyloidogenic isoforms of $\text{A}\beta$ species, in particular $\text{A}\beta_{42}$, the most hydrophobic and insoluble isoform.

The amyloid cascade hypothesis is supported by several AD genetic and environmental risk factors that are associated with enhancing and promoting the aggregation of $\text{A}\beta$, as is the case of the apolipoprotein (ApoE4), which was found, in transgenic mice, to modulate $\text{A}\beta$ production. ApoE is predicted to interact with $\text{A}\beta$, preventing its clearance. The role of risk factors such as excess amounts of free Zinc, Copper, Iron and Aluminum or high cholesterol levels is also connected to ApoE4 haplotype and abnormal $\text{A}\beta$ aggregations. Cholesterol for instance, modulates APP processing and it was found to be correlated with AD. In fact, in animal models elevated levels of cholesterol were associated to promoting $\text{A}\beta$ deposition [14].

An important causative role for other protein in the onset and progression of neurodegeneration has become increasingly evident. Microtubule binding protein tau has been observed abundantly in filamentous form in neurological disorders such as AD. Tau is a low-molecular-weight protein that binds and stabilizes microtubules, promoting microtubule assembly and influencing axonal transport. AD characteristic dysfunction and loss of synaptic connections were initially related with the toxic effects of $\text{A}\beta$. However, recent research places tau as a metabolite able to impact synaptic activity in several ways, suggesting an emerging role for tau at the synapse. Studies show that tau interacts directly with post-synaptic signaling complexes, regulating glutamatergic receptor content in dendritic spines, and influencing targeting and function of synaptic mitochondria. The body of evidence points to the model where tau is in fact the synaptotoxic species in neurodegeneration. Experimental results include mice overexpressing human tau and exhibiting significant synaptic loss; and early trials of tau-targeted immunotherapy reducing tau pathology and also synapse loss [17].

NFTs (neurofibrillary tangles) are intracellular filamentous aggregates of protein tau which, in AD, accumulate in the somatodendritic compartment of neurons due to a initiating pathogenic event that turns soluble tau into its insoluble form. Tau is abnormally phosphorylated in a way that it cannot bind and stabilize microtubules, leading to an increase of free tau levels in the cell body which may aggregate,
resulting in tangle formation and neuronal dysfunction by disruption of normal axonal transport [16].

The amyloid hypothesis suggesting that Aβ precedes and causes tau hyperphosphorylation may be oversimplifying the actual underlying mechanisms of AD. There is evidence suggesting that these two processes may be independent and, in fact, data from autopsied young individuals points to an earlier start of the tau pathological cascade relative to the Aβ deposition [18]. This observation, however, can be interpreted as being a variant of normal aging instead of the beginning of a pathology since it occurs in a relatively high proportion of clinically asymptomatic young individuals.

2.2 Biomarkers of AD

To overcome the barriers emerged when trying to investigate AD, two major research fronts have to be regarded: 1) to update symptomatic assessment protocols and improve sensor technology, whether it is imaging, biospecimen and biosignal data acquisition or neuropsychological evaluations, and 2) to develop more accurate and generalized data analysis procedures, in order to extract more reliable knowledge on the correlation and predictive power of each measurement with the progression of AD. These two research perspectives work tangled with each other. Since the human body is composed by a lot of complex networks modelled by an enormous set of values, from which modern data acquisition techniques can only extract a limited set, it is necessary to know which are the most informative. Therefore, performing a previous knowledge extraction from the available data is a key factor to focus data reading at certain biological and behaviour features.

In this section the current perspective of the known AD key biomarkers, neuropsychological assessment tests and imaging regions of interest are presented, along with a brief review on how this measurements and other risk factors emerged as related to AD, and the usefulness and frequency on which they are incorporated in research studies.

2.2.1 Proteomic Biomarkers

The current understanding of the pathological events occurring in AD cascade can be assessed empirically by examining biofluids. In direct contact with the brain, molecular composition of the cerebrospinal fluid (CSF) is a naturally valid biomarker to analyze in neurodegeneration research, in which has proven to be a valuable incorporation [19].

Abnormally high levels of phosphorylated or total tau protein in CSF are believed to occur after a release from damaged and dying neurons that enclose dystrophic tau neurites and tangles. The large-scale pathological precipitation of abnormally insoluble Aβ peptides leads to a lower concentration of those metabolites in CSF. These two CSF values are quite established as the most promising and informative AD biomarkers [20]. However, further investigation in CSF [15, 21–23] and blood composition levels [24] as revealed, besides the extensively used Aβ42 and tau levels, a significant number of potential biomarkers for AD so far, including biomarkers associated with neuroregulation and oxidative and inflammatory damage, such as isoprostanes, sulphatides, C-reactive protein, 8-hydroxy-deoxyguanine, α-1-antichymotrypsin, interleukin-6-receptor complex, complement C3 and α-1-antitrypsin, as well as biomarkers linked to Aβ aggregation and degradatation, such as Clq, α-2-macroglobulin and ApoE.
A daunting challenge to address when researching for AD therapeutics is indeed the lack of robust biomarkers. Therefore, along with these findings, the incorporation of CSF biomarkers has been exhaustively studied in order to accurately track the progress of the disease and monitor, in a reliable cost-effective way, the impact of new interventions on patients in clinical trials [15].

Several studies using CSF have been performed [25]: Shaw and his colleagues [26], with a data-driven approach, traced an AD CSF profile for t-tau and $\alpha\beta_{1-42}$ which is extensively approached in recent studies as a decisive comparison landmark. ROC analysis by optimizing test accuracy of autopsy-confirmed CSF data showed that the best cutoff points to consider when discriminating AD from NC are 192 pg/ml of $\alpha\beta$ concentration and 93 pg/ml of total tau concentration. The analysis places $\alpha\beta_{1-42}$ as the most sensitive biomarker, with an AUC of 0.913. Furthermore, Samtani et. al [27] developed a model aiming to describe the longitudinal progression of cognitive decline and was able to distinguish two different MCI populations by analysing the respective CSF values: the probable-progressers with $\log(p\text{-tau}_{181p}/\alpha\beta_{1-42})$ greater than -1.86 and the non-progressers otherwise.

### 2.2.2 Neuroimaging Biomarkers

The fact that AD pathology was irrefutably connected with agglomeration of senile plaques, neurofibrillary tangles, neuronal loss, among other factors, led to the creation of a 'Minimum Microscopic Criteria' for the diagnosis of AD, based on the number of senile plaques found in the brain. However, that assessment was only possible in post-mortem autopsies, which consisted in a major obstruction to further understand the disease-specific characteristics. To counter that limitation, Klunk et al. [28] started developing an innovative neuroimaging tracer to track in vivo amyloid-$\beta$ abnormal accumulations in the brain. In 2004, Klunk et. al [29] did the first human tests of a novel prototypical benzo-thiazole amyloid binding agent, termed Pittsburgh compound-B (PIB), indicating that these compounds could bind to amyloid with low nanomolar affinity, infiltrate in the brain in sufficient amounts for imaging with PET, and then be rapidly removed from normal brain tissue. The resulting PET imaging patterns, when compared with normal controls, showed atypical concentrations of PIB in the frontal, parietal, temporal and occipital cortices and in the striatum, as well as no significant difference in PIB retention in areas such as subcortical white matter, pons and cerebellum, known to be relatively unaffected by amyloid deposition. Additionally, PIB attains the interesting property of binding specifically to both extracellular and intracellular forms of $\alpha\beta$ aggregates but not binding to other proteins agglomerations such as NFTs or Lewis bodies, allowing a greater specificity in assessing and discriminating AD from tauopathies and other non-$\alpha\beta$ disorders [30].

Currently, the focus is to develop amyloid imaging $^{18}$F compounds, in order to switch from the use of $^{11}$C PIB, a 20 minute half-life isotope, to the 110 minutes half-life compound and thus obtain greater clinical utility. The potential agents being tested include $^{18}$F-AV-45, also known as florbetapir [30]. Florbetapir administration was shown to be well tolerated by patients and the respective PET images indicated significant discrimination between AD and NC individuals [31].

Besides the recently developed Amyloid-$\beta$ imaging techniques, other neuroimaging techniques are used to study AD and include PET imaging with fludeoxyglucose ($^{18}$F-FDG) tracers and structural magnetic resonance imaging (sMRI), which have been widely used for several years in a variety of diseases, including not only neurodegenerative but also oncologic disorders.
Since it provides a quantitative measure of glucose metabolism, which is directly related to cellular activity, the observation of a progressively decreasing brain uptake of FDG over time in patients with AD is able to reveal neuronal loss and injury in numerous studies. The main brain areas reported to be affected by AD-pathological neurodegeneration using FDG-PET are predominantly the parietal, temporal, frontal and posterior cingulate cortices [32, 33].

Besides conceptually being less related to AD than Amyloid imaging, FDG-PET was shown to be more accurate in monitoring treatment responses of clinically diagnosed patients in later-stages of AD, contrary to PIB-PET, from which is observed a stabilization suggesting a plateau of Aβ accumulation is reached in the later-stages of the disease [34]. Several other studies indicate that FDG-PET may be a better predictor in the last-stages of the disease. In fact, FDG-PET, alongside with anatomical location, is suggested to be a better predictor of cognitive decline than Aβ assessments [35]. The average sensitivity in identifying AD for FDG-PET is 90%, although specificity in differentiating AD from other dementias is lower.[36]

Evaluation of brain structure is also very appealing to neurodegenerative disorders since it is now well understood that the pathological path of such diseases leads to progressive synaptic and neuronal function loss and thus atrophy in a macro scale of target brain structures. This type of structural assessment is made possible by magnetic resonance imaging from which one is able to extract, in high-resolution, metrics such as grey matter, white matter or CSF volumes and, dependent on the applied pre-processing steps, such as brain registration and segmentation, obtain informative metrics for the existing anatomically regions of the brain, such as thickness or volume.

Structural patterns in the brain of AD patients have been extensively investigated for over two decades. Employed as an early diagnosis tool, initial studies focused on medial temporal lobe anatomy, particularly the hippocampus [37], relying on previously well-established pathological knowledge on the characteristic distribution pattern of NFTs in post-mortem AD patients which suggested neurofibrillary pathology beginning in the medial temporal lobe structures and progressing from there, to the paralimbic, basal temporal and other neocortical association areas [18]. Ante-mortem hippocampal measurements from MRI, such as hippocampal volume, rates of brain atrophy and rates of ventricular expansion, correlates with pathological disease indicators at autopsy, such as hippocampal plaque density, neurofibrillary tangle and plaque density, and hippocampal neuron cell counts [37]. Since the first analyses of correlation of MRI patterns with AD pathology, several studies incorporated sMRI to evaluate its potentiality as a biomarker [23, 32, 33, 38–40]. Due to its high-dimensionality and different imaging post-processing methods, classification using MRI is approached from different perspectives. For instance, one can use as features specific measured volumes of regions of interest [37, 41], such as areas known to be affected by AD, obtain features by automated voxel-based analysis [42] or developing new variables that summarize multiple variables or data patterns [43–45].

Similar to FDG-PET, the differentiating power of sMRI may be restricted to later-stages of AD, since it tracks macro changes that occur after several strikes from the pathological cascade. Recently, since the current focus in AD research is to predict potential progression to AD in pre-symptomatic subjects, multiple imaging modalities are being increasingly employed [38, 42, 46, 47].

Regarding other MR techniques, there has been research focusing on functional MRI [23, 33] to obtain
AD-invariant patterns of brain activity. In fMRI, the progressive levels of deoxyhemoglobin are especially appealing to monitor, since its escalations are indirectly related to brain activity. fMRI images, however, have excessive inter- and intra-individual variability to be used as differential diagnostic of dementias [23]. Intrinsic “default modes” have been proposed by observing that some activity variance measured from all areas of the brain was not arose from activations in the resting state [48]. Efforts have been made to unravel changes in functional connectivity in patients with AD while assessing the changes in sub-networks of the default mode network [49].

Bochetta et al. [50] recently analysed the frequency on which key AD biomarkers were being incorporated and its perceived usefulness in research developed in European Alzheimer’s Disease Consortium centers. The survey included biomarkers of amyloid deposition, i.e. amyloid PET and concentration of Aβ and neurodegeneration, including MRI medial temporal atrophy, FDG-PET and CSF tau. The results state that the most frequently used biomarker are medial temporal atrophy assessed by MRI, followed by CSF markers, FDG-PET and amyloid-PET. These findings may be influenced by the availability of special equipment, which will result in a reduced use of more recent procedures, such as amyloid imaging.

2.2.3 Neuropsychological Tests

The cognitive and functional decline that affects the daily lives of the patients with AD and the respective caregivers is now well understood as a consequence of the onset of one of the last periods of the pathological cascade of the disease. Hence, for investigation purposes on AD early stages, neuropsychological measures (NM) have become progressively less important to analyse. However, for most of diagnostic and prognostic algorithms proposed so far, an incorporation of NM is observed and has often been proven to be relevant to the problem and associated to a greater risk of converting to AD.

In order to enable data mining approaches to rely on NM, researchers and physicians have been making great efforts to develop and implement standardized evaluation tests to counter the NM embedded subjectivity (e.g. ADNI protocols). The reality is that the neuropsychological examination is not a mechanical process, the outcome for a certain patient in a specific test depends directly in the skill and judgement of the examiner and the willingness and efforts the patient invests in the test. In most of the AD later stage cases, due to the patient’s dementia severity, the physician is not able to quantitatively evaluate the status of the patient, which results in a general lack of AD late-stages neuropsychological data. Certain practices are recommended when evaluating cognitive skills of AD patients, including conducting tests at approximately the same time to counter possible circadian effects, in a quiet comfortable room and free of distractions. It is important to notice though, that, besides having an associated subjectivity, NM do not require use of expensive equipment, being evaluated by a specialized physician.

According to the recommendations of the National Institute of Aging (NIA) and Alzheimer’s Association (AA) work group [51], the routine diagnosis of AD can still be assessed by purely clinical and neuropsychological criteria. Revising the 1984 criteria for AD dementia [52], McKhan and his workgroup [51], in 2011, summarized, into a unifying criteria (NINCDS-ADRDA criteria), the different behavioural domains that can be affected in dementia, including impairments in: 1) the ability to acquire and remember new information, 2) reasoning and handling of complex tasks, 3) visuospatial abilities, 4) language functions and 5) changes in personality or behaviour. It was stipulated that cognitive or behavioural impairment
required inclusion of at least two of the previous impairments. According to this criteria, a patient can be diagnosed with dementia, in one of these two categories: 1) Amnestic presentation, the most common symptomatic presentation of AD and 2) Non-amnestic presentation, when the patient is impaired in the domains of language, visuospatial and executive function. Along with NINCDS-ADRDA criteria, another gold-standard for clinical diagnosis of AD is the DXM-IV-TR criteria [53].

To assess each cognitive domain, specific neuropsychological tests are employed. The most commonly employed tests in AD aim towards assessing learning and memory, language production and comprehension, executive function, intelligence, visuospatial perception, degree of dementia, degree of depression, daily activity functionality or a mixed combination of these domains. The specific tests are usually chosen depending on the context on which they are applied, the test duration, the degree of effort to the patient and the purpose of the test. The most widely studied and administered NM batteries are the Mini-Mental State Examination (MMSE) and the ADAS-cog (Alzheimer's Disease Assessment Scale - cognitive subscale) [54, 55].

The MMSE was initially created to differentiate organic from functional psychiatric patients, and thus was not recommended to be used, on its own, as a diagnostic tool, being, in other hand, a good instrument to quantitatively estimate severity of cognitive impairment [56]. With a maximum score of 30 points, the MMSE is a set of heterogeneous questions which can be typically arranged in the following categories: orientation to time and place (10 pts), registration of three words (3 pts), attention and calculation (5 pts), recall of three words (3 pts), language (8 pts) and visual construction (1 pt). This battery is currently the estimated most widely employed NM test to monitor cognitive status [54]. Nevertheless, it has raised some concerns about its administration and scoring system, revealing: 1) some lacks in utility when measuring language deficits, 2) limited sensitivity to frontal and subcortical changes, and 3) heterogeneous practices worldwide, with tests being used interchangeably without having the same psychometric properties. Hence, more robust variations have been developed, including the standardized (SMMSE) and the modified MMSE (3-MS) [54].

In other hand, ADAS is an AD-specific test battery to assess general cognitive ability and severity of dysfunction in AD patients, and has been developed due to the perceived lack of monitoring tools in clinical trials. In its classic form [57], ADAS-Cog features 11 items, scaled from 0 (no impairment) to 70 (significant impairment), assessing domains such as memory, language, praxis and orientation. The modified ADAS-Cog 13-item scale [58] was introduced by adding two more cognitive domains to the original ADAS battery, i.e. the number cancellation task and the delayed free recall task, resulting in a more in-depth evaluation of the patient without compromising the duration of administration. Recently, Skinner et al. [55] described a new variant of ADAS-Cog that improves differentiability among MCI patients, by including executive function and functional assessment.

Other multi-domain scales that may be involved in the diagnosis decision process are the Clinical Dementia Rating (CDR), the Disability Assessment Scale for Dementia (DAD) or the Montreal Cognitive Assessment (MoCA).

Since MCI and AD are mainly identified by a significant verbal memory impairment and deficits in episodic memory, several tests specifically target the memory and learning domains. Those assessments usually are used as quick screening tools and include tests such as the logical memory test, immediate
recall, forward and backward digit span from the Wechsler Memory Scale [59], the Rey Auditory Verbal Learning Test [60], the Verbal Paired-Associate Learning (VPAL) and the California Verbal Learning test (CVLT) [61, 62]. Specific NM tests can assess specific cognitive domains that may be affected in the course of the disease, including:

- Executive functions, like problem solving, planning and organizational skills are assessed by tests such as the Trail Making test;
- Language production and comprehension are evaluated by tests like the Boston Naming Test [63];
- Intelligence, described by the ease in performing tests such as the American National Adult Reading Test (ANART) or the Wechsler Adult Intelligence Scale (WAIS) [64];
- Visual and spatial perception are measured by NM batteries such as the Visual Object and Space Perception (VOSP) Battery [65], and the Clock Drawing;
- Daily activities functionality and changes in personality are usually determined by the Functional Activities Questionnaire or the Blessed Dementia Rating Scale.

AD-characteristic cognitive decline and behaviour changes can be the result of an overlap with a depression pathology [66]. Therefore, it is important to account for the bias implied by a possible depressive state. Thus, coupled with dementia and cognitive assessments, patient usually undergo depression tests, such as the Geriatric Depression Scale (GDS), a self-report assessment used specifically to identify depression in the elderly [62].

Making accurate predictions of conversion to individuals in a pre-clinical status using the most easily accessible tests is of major importance. Therefore, some research efforts are focusing on information that can be extracted from these NM and from a simple blood draw [67].

2.3 AD Neuroimaging Initiative

The Alzheimer’s Disease Neuroimaging Initiative (ADNI) is an ongoing American multicenter project designed to find the best predictors for AD and help worldwide research projects to detect more reliably treatment effects in clinical trials. By collecting clinical, imaging, genetic, and biochemical data from patients in various AD stages, the ADNI propelled the discovery and validation of many biomarkers for the early detection and tracking of the disease. The major goals of this study are to validate those as predictors, in order to determine biomarkers for use in clinical trials, and to standardize worldwide data collecting procedures.

Originally, the project was intended to fill the gaps felt in 2000, between academic investigators, pharmaceutical companies, biotech companies and all the parties involved in developing treatments on slowing AD progression. It was felt that the several efforts that were being made were out of synchronization with each other, namely by using different methods in distinct cohorts, and often assessing treatments effects based on unreliable cognitive data, known to be susceptible to change due to symptomatic treatments instead of AD treatments.

Since October 2004, the ADNI has been collecting and studying a diverse set of assessments in
elderly controls, subjects with MCI, and subjects with AD. Since the start, the project went through three study phases: ADNI 1, ADNI GO and ADNI 2.

![Figure 2.3: Timeline of the 3 phases of the ADNI.](image)

ADNI 1, initiated in 2004, was a six-year study funded by $67 million provided by both private and public investments. It included 400 subjects diagnosed with MCI, 200 subjects with early AD and 200 elderly control subjects. In 2009, as ADNI 1 was ending, the project was extended by ADNI GO, adding 200 participants identified as having early MCI (EMCI) to a set of 500 control and MCI patients of the existing cohort. ADNI GO phase extended the study to 2011 (Fig. 2.3), when, funded with another $67 million, ADNI 2 phase started. Assessing approximately 450-500 control and MCI patients from ADNI 1 and 200 EMCI patients from the ADNI GO cohort, ADNI 2 added two new categories - late MCI (LMCI) and significant memory concern (SMC), and introduced 150 normal controls, 150 EMCI patients, 150 LMCI patients and 200 mild AD patients (Table 2.1). ADNI 2 is currently recruiting participants.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Normal</th>
<th>EMCI</th>
<th>MCI</th>
<th>LMCI</th>
<th>AD</th>
<th>NM</th>
<th>MRI</th>
<th>fMRI</th>
<th>DTI</th>
<th>FDG</th>
<th>AV45</th>
<th>PIB</th>
<th>Biosamples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADNI 1</td>
<td>200</td>
<td>400</td>
<td>200</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ADNI GO</td>
<td>190</td>
<td>200</td>
<td></td>
<td>190</td>
<td>200</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ADNI 2</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>200</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Table 2.1:** Number of participants recruited in each phase and types of data.

The ADNI consists on a structure of eight cores, managed by the Administrative Core, directed by Dr Weiner, the principle Investigator. The cores correspond to the Clinical Core, responsible for the recruitment of subjects, electronic data capture (EDC) system and protocols of procedures for ADNI; the MRI and PET Cores, development of standardized imaging methods and all MRI and PET procedures, respectively; the Biomarker Core, responsible for biofluids collection and analysis of biomarkers; the Genetic Core, responsible for genotyping participants; the Neuropathology Core, responsible for autopsy procedures of ADNI participants who die; the Biostatistics Core, responsible for the analysis of statistical models of the generated data; and the Informatics Core, responsible for data-sharing of the data.

Currently, there are 50 ADNI Clinical Trial Sites located across the US and Canada. The establishment of ADNI in the U.S. and the growing necessity for concise cohorts and AD data has led to a stimulation of similar projects in other regions of the world, such as the longitudinal study AIBL in Australia, which counts with 1100 participants with MRI, $^{11}$C PIB and cognitive measures using protocols similar to AD.
In the same manner other efforts can be named, including the more recent Japanese ADNI, with 220 subjects, and the European ADNI, which is enrolling 150 subjects. China and Korea are other regions where several large longitudinal projects are being planned. In general, it is being observed an increase in data-sharing efforts and communication between different research facilities [68].

Table 2.2: Inclusion/Exclusion criteria at screening for new participants in ADNI 2. Detailed information on ADNI study procedures, including excluded medications, exclusion criteria for follow-up patients, and other specifications not shown in this brief table, can be found at http://adni.loni.usc.edu/wp-content/uploads/2008/07/adni2-procedures-manual.pdf. Data are publically available at http://adni.loni.usc.edu/, and were downloaded for this study on May 31, 2015.

ADNI incorporates a vast number of AD key assessments, including sMRI (1.5T and recently 3T), specialized MRI (fMRI, DTI), PET (with PIB, FDG and ¹⁸F-AV-45 tracers), CSF biomarkers and a vast variety of neuropsychological tests. These evaluations are appropriately scheduled according to each patient disease stage, diagnosed at a screening phase. CN, EMCI, LMCI and AD participants will undergo
different study schedules. The longitudinal follow-up of a subject starts with a screening phase where it will be determined the eligibility and collected measures to later use as reference to assess change. Eligibility is determined according to the Inclusion/Exclusion criteria depicted in Table 2.2.

As it can be seen in Table 2.2, the narrowing inclusion criteria leads to some limitations regarding the generalization of the results obtained with the ADNI data. The restricted age of 55 to 90 years leads to some lack of necessary evidence to further corroborate the well-established idea that AD pathology initial phase precedes considerably the onset of the disease, which may be quite before the minimum age required. Furthermore, it can be pointed that the set of assessments performed does not feature possible informative types of data, such as electroencephalogram, magnetoencephalography, magnetic resonance spectroscopy, metabolic and inflammatory markers, and lifestyle information. However, these represent measures that have not yet been demonstrated as good predictors and yield additional burdens to the participants and to the total budget of the study.

Since its inception the ADNI has already provided data for numerous studies, namely some recent work on longitudinal models and prediction of disease onset described in Section 2.4. Researchers all over the world can obtain free access to the data of the ADNI through acceptance of the Data Use Agreement and the submission of an online application form.

ADNI data can be downloaded directly from the website, once the access request is accepted and the user is logged in, allowing the user to acquire the study data, along with raw imaging data and genetic data, including GWAS and gene expression results. The study data consists of the following multiple types of data:

- Information on the clinical assessment, including the diagnosed disease status and neuropsychological battery results;
- Biospecimen results, such as CSF proteins concentration, ApoE genotype, etc;
- Enrollment information, incorporating clinical verification, inclusion/exclusion criteria confirmation, exam dates of the follow-up visits, among other registry information;
- Genotype outcomes, with the respective methodology description;
- Post-processed imaging data, along with acquisition and quality informations;
- Medical History Information, including adverse events, medication history and physical/neurological examination;
- Subject Characteristics, such as the Family History or subject demographics.

Other ways of getting ADNI data include extracting immediate information from the website, which has a user interface to explore and visualize data. Through this user-interface tool it is possible to plot the evolution of a feature across a different one and visualize multiple cohorts with given characteristics and values according to user defined thresholds. The other way to obtain all ADNI data was the one used in this work and consists on

Data can also be accessed through the ADNIMERGE package available in multiple software, in-

---

1For more details on the ADNI complete study schedule consult ADNI web page
2Available upon admission at http://adni.loni.usc.edu
3More detailed description of the package is available in http://adni.bitbucket.org/
cluding the version used in this thesis implemented in R\(^4\), and other versions adapted to SAS, Stata and SPSS. This package includes and daily updates the most commonly used demographic, clinical, MRI and PET data variables.

2.4 Related Work

To improve the methods applied to the problems of AD diagnosis and prognosis prediction and to gradually build a knowledge basis in which future work can rely on, does not mean solely to optimize the output classifier’s metrics. It also means to acquire new knowledge about the disease-specific cascade of biological and symptomatic event and to model the underlying complex disease mechanism as simple as possible.

The related work presented here will describe in detail the state of the art methods which are being applied recently on data from the ADNI’s database. According to the two major underlying contributions/goals of this work, we group the relative work in the descriptive studies (Section 2.4.1 - Longitudinal Models) and the classification approaches (Section 2.4.2 - Data Mining). The detailed information about the specific distributions in the datasets used in the present studies and the different employed features are depicted in Appendix A, in Tables A.1 and A.2, respectively.

2.4.1 Longitudinal Models

The current view on the way AD progresses differs from the model accepted 2 or 3 decades ago, in which we assumed individuals who had AD pathological changes, had consequently dementia and individuals who did not have those changes, thus did not had dementia. Currently, AD’s pathological cascade is perceived as a gradual deterioration coupled with a gradual clinical decline. However, the way on which the clinical component relates to the pathological aspect of the disease is still unclear. Biomarkers association with AD is being unravelled but there is a necessity to develop a coherent time-dependent ordering of biomarkers in order to better understand the course of disease and to more efficiently introduce and interpret biomarkers in future research. For this purpose, Jack et al. [3, 4] proposed a conceptual model of the dynamics associated with the pathological cascade in AD (Fig. 2.4). The model hypothesises that the key AD biomarkers deteriorate in a temporally ordered manner, in which CSF A\(_\beta\) and amyloid PET are precedents of CSF tau and FDG PET, being followed by structural MRI and finally by the clinical symptoms. The order associated to AD does not mean that early biomarkers have to change completely before the next one starts progressing, rather it indicates a sequence for the periods where each biomarker will experience the greatest changes. The specific period of onset of clinical symptoms is modelled with an inter-subject variability mediated by non-biomarker factors such as brain resiliency and cognitive reserve.

Experimental tests to empirically support the basic foundation of this model are mainly made by two perspectives: 1) investigating the real shape of the progression curves or 2) estimating the AD-characteristic biomarker sequence and variance. Evidence accumulated clearly supports the temporal ordering framework considered. In other hand, studies regarding the shape of the curves generally con-

\(^4\)Open-source software platform for statistical data analysis, implemented within the system. Available at http://www.r-project.org
clude that the trajectory of non-Aβ variables may not necessarily be parallel curves to the slope representing the accumulation of amyloid plaques. The stabilization at later-stages usually is not observed. FDG PET and MRI atrophy suggest a continuation of declining in the dementia phase of the disease.

To further accumulate empirical evidence and develop data-driven model, some concerns have to be regarded. First, research must always consider the fact that AD data is often contaminated with non-AD pathology, whether if the patient has a mixed pathologies progressing simultaneously, such as vascular disorders, Lewy bodies, TDP-43 inclusions, or if it is the case of a misdiagnosis, due to the overlapped symptoms and signs of different neurodegenerative disorders, that can mislead the physician's decision. Secondly, data-driven models based on this model have to define an appropriate unit for the time axis, which has proven to be a challenge, since the pre-clinical phase of the progression, consisting in maybe more than half of the total period of disease, is imprecisely depicted in the data.

![Figure 2.4: Model of biomarker deterioration dynamics](image)

**Figure 2.4:** Model of biomarker deterioration dynamics [4]. The model postulates a logistic increase of biomarker severity across time, with a distinct maximum slope period for each main aspect of the disease. Aβ progression, represented by CSF Aβ42 (purple) and amyloid imaging (red), initiates the progressive cascade and is followed by elevated CSF tau (blue). Neurodegeneration (orange) is tracked by structural and brain activity markers, like MRI and FDG PET. The final cognitive declining state is reached depending on subject-specific aspects, such as cognitive reserve or lifestyle, which give rise to a green filled area in the model, with a low and high risk boarders. The vertical black line denotes a data snapshot at time T. A single snapshot values are considered as the projection of the intersects of the vertical line with the biomarkers curves.

### 2.4.1.A Long-term Progression Curves with Short-term Data

Recently, Donohue et al. [5] proposed a method to recover long-term disease trends from short-term data, motivated by the fact that AD progression starts long before the onset of symptoms. They created an algorithm to estimate population curves of biomarkers severity for decades of AD, by incorporating information from patients in distinct stages of the disease.

The main idea is to expand data time scale, which is limited to short-term follow-up visits (usually not greater than 6 years), into a long-term time scale that can reach a range of 20 or more years. For that purpose, each subject will be shifted forward or backward in the long-term timeline according to
the performance in the estimated progression curves across the panel of variables. At the end, the algorithm outputs the estimated biomarker-specific long-term curves, along with the modelled subject-specific random-effects and time-shifts. The variables incorporated in the model include data from multiple modalities, including imaging, NM and CSF measures (Table A.2).

To test the algorithm and for demonstration purposes only, Donohue and his colleagues simulated reconstruction of long-term trends after rearranging generated long-term data into short-term data. The data was extracted from logistic, linear and quadratic functions and divided randomly into a number of sets, simulating aggregates of data from different subjects. To transform into short-term data, i.e. simulate real data, the data for each subject was centered in time zero. When applying the algorithm to that short-term data, they were able to reconstruct long-term curves with good fidelity to the true generating functions. Testing in different-shape curves highlighted the versatility of this approach, which, by being nonparametric allows a reliable curve estimation without pre-specifying any parametric families.

It was used data from the ADNI, namely from subjects with some evidence of AD pathology, i.e. a cohort who revealed abnormal values defined by published standards - lower accumulation of amyloid in the brain (Aβ < 192 pg/mL), PIB PET SUVR greater than 1.5 and AV-45 PET SUVR greater than 1.1 - and also a cohort with subjects with at least one ApoE ε4 allele (further detailed in Table A.1).

The analysis of the resulting panel of curves suggests that amyloid PET (AV-45 and PIB) may be the initial signal of the AD pathophysiological cascade, followed by abnormal values in CSF measurements. Glucose metabolism (FDG-PET), learning and cognition dysfunction and hippocampus atrophy all occur in close succession. Functional difficulties, as expected, were estimated to be the last signal to become apparent.

When closely observing the estimated curves, it becomes evident some limitations regarding the data available. For instance, a lack on follow-up visits for subjects with later stage dementia leads to a poor estimation of the final stages of the long-term curve, which may be the cause for FAQ scores to follow a parabolic trajectory, without demonstrating a final plateau of a sigmoid. The high heterogeneity and between-subject variability, well accentuate in CSF measures, tends to flatten the mean trajectory, and is planned to be investigated in future works by incorporating covariates into the model and adopting hierarchical random effects. Regarding the comparison between cohorts (amyloid+ and amyloid-, ApoE ε4+ and ApoE ε4-), the resulted differences are difficult to interpret, since it is not certain that non-AD pathology is truly depicted in the cohorts without AD signs, such as the amyloid- or ApoE ε4-. However, interpreted to some extent, the amyloid- trajectory shows us some divergent biomarker signature, such as normal CSF levels and less sharp hippocampal atrophy and ventricular expansion rates.

The most efficient measures of disease progression, in terms of the signal-to-noise ratio (ratio of the first derivatives of curves and the residual standard deviation), were provided by structural MRI, which dominates the other measures across the whole timespan. In general, the observations are in agreement with the current idea of biomarkers progression in AD, hypothesized in the model of Jack et al. [3, 4].

2.4.1.B Alzheimer’s Disease Event Ordering

Young et al. [69] recently applied the event-based method (EBM) [70] on ADNI data to investigate the sequence of key biomarkers events (moments on which biomarker levels transit from normal to abnormal
range) and to provide a patient staging system. The model constructs a Bayesian framework for the estimation of an AD-characteristic event sequence on a particular data set. This approach assumption of a universal ordering common to all patients within the disease cohort is a major oversimplification of reality. However, it can be consistent enough to distinguish AD from other diseases.

The probabilistic generative formulation is based on the likelihood of a feature value to belong to the pre- or post-event cohort, i.e. whether if it is more likely for that particular feature value to represent a normal or abnormal situation. In a previous paper [70], where EBM was applied to familial AD data, the pre-event likelihood distribution was obtained by fitting the control cohort data to a Gaussian curve and then fitting the remaining data to the assumed-uniform post-event likelihood distribution. However, for ADNI data is quite unreliable to assume a cohort of control subjects because a great fraction of the participants diagnosed as CN are in fact in a pre-symptomatic AD stage. Nevertheless, for each feature distribution, it is expected to be noted two peaks corresponding to the distributions of subjects with normal and abnormal values, i.e. values before and after that feature’s event. Taking this into account, Young et al. proposed fitting a mixture of two Gaussian distributions to the data, with a simple constraint dictating that the respective standard deviations had to be smaller than the ones observed in the CN and AD cohorts, to avoid highly overlapped models.

Once obtained the distributions for the likelihoods of a value to be on a pre- or post-event situation and assuming that all variables are independent and that there is an equally a priori probability of a subject to be in any position of the sequence, a Bayesian formulation is formulated, defining the posterior distribution of the possible sequences knowing the data.

Since the posterior distribution is analytically intractable, it was necessary the implementation of a Markov Chain Monte Carlo algorithm, a sampling technique. With a greedy ascendant pre-step and sufficient MCMC repetitions the procedure acquires a sequence set quite varied while relatively close to the maximum-likelihood solution. This approach and additional data bootstrapping allowed the estimation not only of the maximum likelihood event sequence (characteristic event sequence) but also of the respective positional variance.

The results suggested that CSF measures are the first to be affected, followed by the rates of brain atrophy, then cognitive test scores, and finally regional brain volumes. Similar to Donohue and colleagues work [5], a comparison between using the whole population dataset and using probable AD cohorts (amyloid+ and APOE ε4+) was performed, and the results did not obtain significant differences, except a reordering within the CSF biomarkers events. Additionally, the separability and predictive power when using the model’s staging system was tested. For each patient, the stage that maximizes the probability of the data given the characteristic event sequence was computed. Since the majority of CN were positioned in stage 0, estimated that no event had occurred, and the patients with diagnosed AD were assessed as being in the later stages, using the patient staging system to classify between CN and AD individuals resulted in a balanced accuracy of 99%.

Regarding the prognosis predictive power, the model stage was found to be a significant risk factor for conversion using Cox proportional hazards models, achieving 77% of balanced accuracy and AUC of 0.78 over 3 years in predicting conversion from MCI to AD. Furthermore, it is important to notice the great potential of this approach to quantify the uncertainty of the model and calculate the overall likelihood of
a specific patient staging on the event sequence. Once available, those results may be important points
to incorporate in a learning method and also to provide to physicians, who will be able to interpret and
understand the certainty on which the algorithm outputted the results.

For the same problem, Huang et al. [71] presents the AD Probabilistic Cascades model (ALPACA), a
generative model similar but more generalized than the EBM. Contrary to EBM, the proposed algorithm
does not assume multiple sets of measurements from a single patient as independent, introducing the
"snapshot set", another latent variable besides event ordering.

“Snapshot” is the term given to the assessment set of a patient in a specific date and "snapshot set"
is the set of positions where the snapshots fit the sequence (e.g. \{2,3\} as a snapshot set means that the
periods when the two snapshots of that patient occurred correspond to the moments after event 2 and
3, respectively). With this new variable, Huang and his colleagues introduce a new degree of freedom
to reduce residual error, which otherwise would be completely incorporated in the event sequence esti-
mation, as is the case for the EBM. To account for the intractability of the problem, they tried to apply a
Gibbs sampling approximation to generate sequences from the implicit Mallow distribution and snapshot
sets from an uniform distribution. However, since Gibbs sampling was not trivially applied to this problem,
they used an existing method from which independent samples can be efficiently and exactly drawn and,
for snapshot sets, they presented a dynamic programming algorithm based on a grid representation of a
hidden Markov mode, with running time much lower than the exhaustive search.

Contrary to Young et al. [69], this model fits the likelihood distributions of pre- and post- event to
healthy control data and to AD subjects data, respectively. The algorithm was tested with synthetic data
experiments for various data generating settings, and was then applied to a cohort with 347 subjects from
the ADNI dataset (Table A.1) using seven key biomarkers (Table A.2).

The results state that the optimal central ordering is ADAS, followed by hippocampal volume, Aβ,
hippocampal atrophy rate, tau levels, brain atrophy rate and brain volume, which, besides ADAS, closely
tracks the hypothesized pathological cascade. It was also compared the inference using EBM in the used
dataset, which resulted in an event sequence with some conceptual incoherence, such as placing Aβ and
tau events in the later stages of progression, and with a quite higher Bayesian information criterion.

2.4.2 Data Mining with ADNI Data

As we will see in this section, there is a variety of approaches to the problem addressed in this
thesis. In general, the classification studies can be summarized by the methods used regarding instance
selection, feature extraction, feature selection and the machine learning training technique. Table 2.3
summarizes the techniques applied in the six studies presented in detail in this section. Also in the
context of this thesis, regarding improvement of data mining approaches performance, it can be found a
great amount of other studies addressing the problems of diagnosis classification, prognosis prediction
and cognitive decline prediction, but which will not be presented here [43, 72–81].
Table 2.3: Techniques applied regarding different methodology aspects of the studies presented: Ewers et al.[82], Zhang et al.[83], Cui et al.[84], Cheng et al.[85], Ye et al.[86] and Barnes et al.[87].

Ewers et al. [82] constructed a logistic regression model for classification between AD and Normal subjects (NL) and used it for prediction of the conversion of MCI to AD, by testing the model in amnestic MCI data. The data was acquired from ADNI’s database, resulting in CSF, MRI and neuropsychological data set from 182 subjects (Tables A.1 and A.2). The constructed model was the one more frequently constructed from 1000 trainings with random-split resampling. The diagnosis classification of AD vs NL yielded 94.8% and 86.7% of accuracy with and without inputting NM features respectively. Additionally, it was incorporated the model LR$_{TAA}$, a established regression formula based upon CSF concentrations of t-tau, A$\beta_{1-42}$ and number of ApoE 4 alleles [26]. With this new feature, the results were 95.2% and 91.1% of accuracy with and without NM features, respectively.

When testing the model with MCI dataset in order to predict conversion of MCI to AD, the resulting accuracies were 64.1% and 62.5% without LR$_{TAA}$ and 68.0% and 64.8% with LR$_{TAA}$, with and without NM respectively. With MCI instances, time to conversion to AD was tested using Cox regression analysis. They also assessed the predictive value of different combinations of up to 4 variables when training with MCI data to predict conversion. The highest accuracy result, 76.3%, was obtained when training with Trail Making Test B (TMT-B), right Hippocampus volume, CSF p-tau$_{181p}$/A$\beta_{1-42}$ ratio and age. Single-variable training resulted in 68.5%, with the top-ranked non-NM predictor being the right Entorhinal cortex thickness and 64.6%, with the top-ranked NM predictor, TMT-B.

Zhang et al. [83] has used multi-modal imaging data from MCI patients to construct a multi-kernel SVM model for cognitive change regression and MCI-AD conversion classification. The data, composed of MRI, PET and neuropsychological measures (Table A.2, Page A-3) from 5 different time-points, was acquired from the ADNI’s database and was narrowed down to 88 MCI subjects who checked quality requirements (Table A.1, Page A-2).

Imaging feature selection was performed taking into account the longitudinal characteristic of the features, by applying feature selection on multiple time-points, unlike the existing single-time-point based feature selection methods. This was performed by optimizing a regression with a regularization term.
to analyse features as a group across all time points. The method allowed to consider the longitudinal predictive value of the different variables depending on their evolution over time, as well as its individual value. The resulted output is a sparse weight matrix for the different data modalities and for the various time-points from which they extracted the non-null features. The selected top feature regions, include hippocampal, amygdale, entorhinal cortex, and para-hippocampal regions.

In addition to feature selection, it was estimated features based on higher-order features. Solving a higher-order linear equation for each subject and for each selected brain region, they managed to obtain new features such as thinning speed, from thickness-based features.

The training was performed using a multi-kernel SVM. It was constructed an individual kernel for each feature modality and then optimized a mixed kernel based on the linear combination of all individual kernels. This method was trained to predict future clinical scores and future conversion. For the former, the methods predicted MMSE and ADAS-Cog scores, 24 months after baseline, using all data acquired at previous time points. For comparison, they tested, with the same data, CONCAT and Ensemble, other multimodal regression methods, which mainly differ in the way longitudinal data is concatenated. The improvements obtained when including more time-points were poorer using the other approaches than using the proposed algorithm. Pearson’s correlation coefficients and root mean square errors were calculated for each method and when incorporating more time-points, this method’s correlation rose from 0.66 to 0.79, outperforming the other two methods, which went from 0.67 to 0.69 (Ensemble) and from 0.64 to 0.7 (CONCAT).

Regarding prediction of conversion from MCI to AD, the proposed method obtained an accuracy of 78.4%, sensibility of 79%, specificity of 78% and an AUC of 0.768 when using both baseline and longitudinal data, which outperforms both approaches using CONCAT (70.5% acc, AUC 0.742) and Ensemble (65.9% acc, AUC 0.706). This further validates the conclusion that incorporating complementary data domains and including multiple time-points to engage differences in longitudinal feature signatures leads to improvements in the prediction of future clinical scores and future conversions of MCI subjects.

Cui et al. [84] constructed a SVM classifier of AD vs NC and has used it to classify MCI data and predict conversion to AD. The data, composed of MRI, CSF and NM (Table A.2), was acquired from the ADNI’s database and was narrowed down to 350 subjects (Table A.1) who checked quality requirements and had at least 24 month follow-up evaluations.

The methods used can be summarized in two steps: feature selection and SVM training. Selection of MRI and CSF variables was performed using a minimum redundancy and maximum relevance filter to obtain a feature ranking. From that ranking, it was applied a wrapper method where the subsets tested were the results of incrementally adding features from the highest to the lowest on the ranking. Optimal features were selected when the highest AUC was obtained. They performed 20 repetitions with 10-fold cross validation and computed selection frequency for each each feature. The most discriminative feature subset was identified as the subset of features with over 50% selection frequency. This resulted in narrowing the CSF and MRI feature set, from 5 to 2 and 323 to 7, respectively. For selected MRI features, the results were, ordering by selection frequency, Left Entorhinal Cortex, Right Middle Temporal Gyrus, Right and Left Hippocampus, Right Inferior Parietal Cortex, Left Retrosplenial Cortex and Left
Middle Temporal Gyrus. Regarding the CSF features selected, it was narrowed down to t-tau/Aβ_1-42 and p-tau/181p/Aβ_1-42 ratios. NM data was selected in a different way, since it was assessed as being highly separable between AD and NC and resulting in selection of the top 1 feature every repetition. Therefore, it was not used the described wrapper method. Instead, the most discriminate subset, composed by 5 features, was selected from the measures with a correlation coefficient above 0.3 and classification AUC above 0.95. That selected features included FAQ, LM delayed recall, LM immediate recall, AVLT delayed recall and AVLT trials 1-5.

SVM classification yielded the best results when using all the different types (67.13% prediction accuracy, AUC 0.796). Testing the classifier when using various combinations of data modalities led to the conclusion that NM were the more predictive data. NM features were incorporated in all top combinations, and yielded 65% of accuracy and AUC of 0.76 when used individually, which overcame the results obtained when using MRI plus CSF (58.74% prediction accuracy, AUC 0.673).

Additionally, the correlation between the distance from a subject to the SVM hyperplane and the time-to-conversion (ttc) was analysed. The results obtained were only statistically significant when comparing the two cohorts with ttc greater and less than 12 months, which is disappointing since prediction of conversion to dementia in less than 1 year is already quite directly estimated by the physician.

Cheng et al. [85] has constructed a semi-supervised multi-modal relevance vector regression (SM-RVR) model for cognitive change estimation. RVR is a sparse kernel method formulated in a Bayesian framework, which finds the optimal weight vector that approximates a linear combination of basis kernel functions to the output variable. Here the model is extended to consider multimodal data using linear combinations of kernel functions for each data domain. The data, composed of MRI, FDG-PET, CSF and neuropsychological measures (Table A.2, Page A.2), was acquired from ADNI’s database and was narrowed down to 202 subjects (Table A.1, Page A-2).

Different combinations of the three data modalities were used to train the model. It was also investigated the performance of the method when applying an a priori semi-supervised instance selection to choose the most informative MCI subjects to include in the training set. This was performed by estimating a new cognitive score for each MCI subject from the mean cognitive scores of the k nearest AD/Normal subjects across all variables. With the new estimated cognitive score, MCI subjects are individually included in the training of an M-RVR model and the top at minimizing the regression’s RMSE are picked.

When using semi-supervised instance selection, the best performance included all three modalities and obtained a root mean square error (RMSE) of 1.92 and a correlation coefficient (corr) of 0.80 for MMSE scores, and a RMSE of 4.45 and a corr of 0.78 for ADAS-Cog scores. The closest results were obtained using MRI and CSF, with a RMSE of 2.03 and a corr of 0.77 for MMSE scores, and a RMSE of 4.98 and a corr of 0.74 for ADAS-Cog scores. The poorest results were obtained when using only CSF data to train the model. It was observed that SM-RVR consistently outperforms two different versions of M-RVR on each performance measures, clarifying the importance of pre-excluding less informative MCI subjects.

The proposed SM-RVR method was also used for predicting the future conversion of MCI subjects. The SM-RVR was trained using AD, NC and MCI data with new estimated cognitive scores. The achieved
prediction accuracy was 69.4% (with sensitivity of 72.4% and specificity of 67.8%). In contrast, when training a M-RVR considering the training MCI converters as AD and the non-converters as NC, it is obtained a lower prediction accuracy of 66.4% (with 69.3% sensitivity and 63.7% specificity). M-RVR trained with only the AD and NC as training set, achieved a classification accuracy of 58.4% (with 64.1% sensitivity and 54.1% specificity).

Ye et al. [86] constructed a SVM model using logistic regression stability selection for prediction of conversion from MCI to AD. The data was acquired from the ADNI's database and was narrowed down to 319 MCI subjects (Tables A.1 and A.2). Feature selection was performed by fitting a sparse logistic regression with norm regularization, and selecting the non-null weighted features. By repeating the procedure for different datasets (bootstrapping) and for various regularization parameters $\lambda$, the procedure is statistically more robust and stabilized, thus the term "stability selection".

Compared with a feature selection based on a t-test ranking, the resulted prediction AUC, with stability selection, improved from 0.79 to 0.86. The top 15 features found by stability selection order FAQ as the most important feature, followed by APOE$\varepsilon$4, ADAS-Cog sub-scores, Logical Memory Delay, Volume of Left Hippocampus and other volumetric measures. Memory and orientation tests from ADAS-cog sub-scores are in the top six and most of the MRI features selected are volumes known to be reduced in AD.

Comparisons between using different features types highlighted the benefits of the combination of demographic, genetic and cognitive measurements. For instance, it was observed that using MRI alone (AUC 0.72) is outperformed by the use of multi-modal combination (AUC 0.86). Another interesting conclusion was that, by including CSF markers, the prediction yielded lower AUC, making CSF poor predictors in this framework, which goes against the conclusions made by the LR$_{TAA}$ model [26].

Barnes et al. [87] constructed a point score-based model for prediction of conversion from amnestic MCI, outputting the risk level of conversion (high, medium or low). The model, after a series of feature selection steps based on Cox proportional hazard analysis, ranks the final features with a score (0-9) according to the coefficients of a final Cox model. The data was acquired from the ADNI's database and was narrowed down to 382 amnestic MCI subjects (Table A.1) who had at least one follow-up visit. Additionally, it was made a parallel experiment using two sub-datasets, resulting from separating the 382 subjects into subjects with age lower and higher than the mean age (75 years). The method uses various data domains (Table A.2) excluding CSF and PET variables since they are scarce in ADNI database.

For each of the different data domains, the proposed method starts by selecting the features that were associated with conversion to AD ($p<0.2$) through a Cox analysis. For the continuous variables, it was discretized the possible outputs into quartiles and, when available, into clinically meaningful categories. In the second step, the selected features of the most predictive subset, regardless of domain, compete against each other in a final Cox analysis. The model uses the features significantly associated with conversion ($p<0.05$) as independent predictors, by assigning them a point-score proportional with the respective final Cox analysis coefficient. The outcome is a table mapping each quartile or interval of selected features values with a point score ranging from 0 to 3. The resulted table, consisting in the key predictors, includes FAQ (0-3 points), MRI middle temporal cortical thinning (0-1 point), MRI hippocampal...
subcortical volume (0-1 point), ADAS-cog (0-3 points) and the Clock Test (0-1 point). To each subject is assigned the total score resulting from the sum of the sub-scores from each of the correspondent features values. The total point scores of the subjects ranged from 0 to 9, with a mean of 4.8.

The resulting Harrell's c statistic of the predictor was 0.78 (95% CI 0.75-0.81), and 0.74, when applying bootstrapping. These values are relatively lower than the 0.78 (with NM and MRI) and 0.8 (with NM, MRI and CS) obtained in Cui et al. [84] and also lower than the 0.86 obtained in Ye et al. [86]. For the sub-dataset of subjects with less and more than 75 years, the model predicted conversion with a Harrell's c of 0.8 and 0.78, respectively.
Contents

3.1 Proposed Approach .................................................. 32
3.2 Data Preprocessing .................................................. 33
3.3 Computing Learning Examples .................................... 42
3.4 GRACE Models ...................................................... 44
3.5 Classification Methodology ......................................... 50
3.1 Proposed Approach

In this thesis, we propose the extraction of interpretable features from longitudinal models of progression of biomarker severity along with the incorporation of those features in a data mining routine to predict conversion of MCI to AD within 3 years. Furthermore, we provide a methodology (Fig. 3.1) with the purpose of yielding important knowledge to support physician’s decision making and facilitate upcoming research in this field. This approach describes the way different biomarkers progress over time and estimates the stage of a patient in the long-term disease sequence. To account for the longitudinal changes within-subject, each instance on the training set will correspond not to a specific patient, but to each visit of each patient (snapshot).

The models applied in this thesis were based on the algorithm from the work of Donohue et al. [5]. This algorithm, termed GRowth model by Alternating Conditional Expectation (GRACE), iteratively estimates biomarker progression curves while transforming the data by shifting each patient data on the long-term timeline. From this algorithm, the estimated long-term progression curves cover more than two decades of disease progression, using data with short-term follow-up periods, typically not greater than six years. Alternating with fitting the model, each subject is assigned a temporal shift $\gamma$, depending on the performance of the biomarkers values on the estimated curves. The point on the long-term timeline which best fits each subject’s values can be interpreted as the disease stage of that patient, relative to the other subjects in the dataset used to construct the curves. A direct time span to the onset of symptoms cannot be extracted, since the temporal scale is not anchored to a reference point.

The features extracted from the model include the subject-specific temporal shift $\gamma$ along with the estimated disease stage $t_\gamma$ of each snapshot. The $\gamma$ correspond to the relative deviation that each patient has undergone from its initial point on the original dataset to the time point assigned after the completion of the algorithm. The $t_\gamma$ corresponds to the absolute value on the constructed long-term time axis after applying the shift $\gamma$, i.e. it is the estimated disease stage for that patient. Despite the interpretability of the proposed features not being completely obvious, since we do not anchored the timeline to a specific real event moment, introducing these variables in a classifier is expected to yield improved results, since we are adding greatly separable features.

The data was obtained from the ADNI database through the R package ADNIMERGE.1. This package was created by Michael Donohue and used in his work [5]. We performed similar pre-processing techniques to prepare the data and reproduce the model results. However, extra pre-processing steps were incorporated in order to include data information ignored in Donohue’s work, namely implementing routines to incorporate visit codes and clinical diagnosis changes along with a correction of inconsistencies found in those variables.

The experimental pipeline culminates in two distinct set of results, resulting from the training set with the 19 baseline (BL) features, which include key imaging, genetic, CSF and neuropsychological assessments, and the training set with the 19 BL features merged with the model-extracted features $\gamma$ and $t_\gamma$.

1Available upon admission at http://adni.loni.usc.edu More detailed description of the package is available in http://adni.bitbucket.org/
which we termed GRACE+. The aim of this thesis is to assess the predictor changes resulting from the dataset GRACE+, i.e. evaluate the results of including GRACE model variables $\gamma$ and $t+\gamma$.

The target predictors are learned from 75% of the total dataset and combine different learning methods and data processing techniques. The classifiers used (implemented in WEKA [88]) are: C4.5 Decision Trees, K-Nearest Neighbour, Logistic Regression, Naïve Bayes, Random Forests and Support Vector Machines (with Gaussian and Polynomial Kernels). Data preprocessing was also tested by combining Feature Selection, Missing Values Imputation and Synthetic Oversampling (SMOTE). The training methodology was developed on a 10-fold Cross-validation (10-fold CV) framework. The classifier and pre-processing techniques that optimize the results of the 10-fold CV in the training data (75% of total dataset) are picked to represent each tested dataset (BL and GRACE+).

![Proposed methodology pipeline](image)

**Figure 3.1:** Proposed methodology pipeline. The work proposed in this thesis is built upon the framework of [5]: The “Input” represents the data, stored in various data files, obtained from the ADNIMERGE R package, and the curve estimation algorithm, available in the GRACE R package [89]. The “Methodology” layer represents the different steps of this work: 1) Pre-processing, including the methods employed in the previous work and also the corrections implemented in this thesis; 2) Construction of GRACE models using data from all diagnosis categories; 3) Creation of learning examples, by aggregating the data into instances and assigning classes, with and without GRACE features; and finally 4) Classification routine. The resulting “Output” of the proposed pipeline are the progression curves along with the prediction metrics, using the default feature set (Baseline) or the augmented feature set (GRACE+).

### 3.2 Data Preprocessing

Real-world data invariably incorporates a given amount of measurement noise, and is often incomplete and inconsistent. In the dataset used in this thesis (described in Section 3.2.1), human errors in data insertion, AD misdiagnosis, lack of measurement quality, among other factors may be the cause of some missing values and incoherent data. As a result we apply data filter techniques with the purpose of data cleaning, integration and transformation or size reduction, in order to try to obtain better classification results. It is important to notice that even if a given methodology results in lower accuracy or AUC score, it may be compensated by an increase on efficiency or simplicity of the generated models. The next
subsections present the routines implemented to correct and adapt data to the problems of classification not addressed by Donohue et al. [5]. The preprocessing techniques that were already incorporated in the GRACE previous work are mainly related to preparation of data to be modelled and will be presented in the Section 3.4, along with the foundation of the model. In this section, we will discuss in detail the inconsistencies found in the data and how we approached them.

To accomplish the target goal, we need extra information relatively to the existing implementation of GRACE preprocessing. Indeed, one problem which is not explored by the GRACE algorithm is the conversion of the dataset into data instances for a classifier. The GRACE algorithm models each variable as depending solely in its value and the respective exam date, and consequently does not need to aggregate multiple variables into one single instance. However, in this thesis, in order to apply a classifier to the data, a preprocessing routine is required (Fig. 3.2) to incorporate two variables: one variable to guide the aggregation of assessments into instances (Section 3.2.2) and one to characterize instances with a class concerning the progression to AD (Section 3.2.3).

![Figure 3.2: Preprocessing routine implemented to extract from the ADNI files, clean and correct two types of information regarding the way instances are labelled and the progression of the diagnosis status. From the ADNI files, the assessment values are extracted as will be described in Section 3.2.1, along with the files Registry and DXSum, which carry the data necessary for the steps described in this section. The two datasets are merged and then is proceeded with a correction to clean inconsistencies and fill some missing values in those variables (represented by orange in the merged file).](image)

### 3.2.1 Dataset Statistical Description

The dataset used in this work is composed by data from normal elderly, MCI subjects and patients with diagnosed AD, recruited by the ADNI (described in Section 2.3) and includes data from all phases of the study, which covers participants from January 2005 to January 2015. It comprises longitudinal demographic information, neuropsychological evaluations, neuroimaging, ApoE genotyping and CSF biomarkers, along with two files with complementary data regarding registry information (Registry file), including visit codes and exam dates, and regarding the output of sporadic diagnostic clinical reviews (DXSum file). The Registry file possesses data of each patient’s visit admission, such as the registry date and the ADNI cohort associated. The DXSum summarizes diagnosis information and comprises the variable DXChange where the outcome of a diagnosis clinical review is expressed.

The NM are represented by the results of the RAVLT, ADAS 13, CDR and MMSE, along with func-
tionality assessment, determined in the FAQ results. Genetic information is summarized by the ApoE ε4 number of alleles. MRI-extracted features include 4 of the more relevant areas affected by AD, consisting in the Hippocampus, Entorhinal, Ventricles and Whole Brain volumes. PET values from three different tracers, AV-45, PIB and FDG were summarized into single values by combining PET intensities from different regions of the brain. CSF proteomics include the concentrations of proteins tau, p-tau and Aβ. The exact transformation performed from file variables to this features is summarized in Table 3.1.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Feature</th>
<th>File</th>
<th>File equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification</td>
<td>Visit Code</td>
<td>registry</td>
<td>VISCODE</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>DXChange</td>
<td>dxsum</td>
<td>DXCHANGE</td>
</tr>
<tr>
<td>Demographic</td>
<td>Age (baseline)</td>
<td>arm, ptdemog</td>
<td>AGE</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td></td>
<td>PTGENDER</td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td></td>
<td>PTEUOCAT</td>
</tr>
<tr>
<td>NM</td>
<td>ADAS13</td>
<td>adas</td>
<td>ADAS 13 Total Score</td>
</tr>
<tr>
<td></td>
<td>FAQ</td>
<td>faq</td>
<td>FAQ Total Score</td>
</tr>
<tr>
<td></td>
<td>MMSE</td>
<td>mmse</td>
<td>MMSE Total Score</td>
</tr>
<tr>
<td></td>
<td>CDR</td>
<td>cdi</td>
<td>Memory + Orientation + Judgement and Problem Solving + Community Affairs + Home and Hobbies + Personal Care Scores from CDR</td>
</tr>
<tr>
<td></td>
<td>RAVLT</td>
<td>neurobat</td>
<td>Trial 5 Total Score – Trial 1 Total Score</td>
</tr>
<tr>
<td>MRI</td>
<td>Hippocampus</td>
<td>ucsfssx, ucsfssx51</td>
<td>Hipocampus Bilateral MRI Volumes x 1000</td>
</tr>
<tr>
<td></td>
<td>Ventricles</td>
<td></td>
<td>Ventrices Bilateral MRI Volumes ICV</td>
</tr>
<tr>
<td></td>
<td>Entorhinal</td>
<td></td>
<td>Hipocampus Bilateral MRI Volumes</td>
</tr>
<tr>
<td></td>
<td>Whole Brain</td>
<td></td>
<td>WM hypointensities + Cerebellum cortex + Cerebellum WN + Thalamus + Caudate + Putamen + Pallidum + Hippocampus + Amygdala + Ventral Diencephalon Bilateral MRI Volumes + Accumbens Bilateral MRI Area</td>
</tr>
<tr>
<td>PET</td>
<td>PIB PET</td>
<td>pibpetsuvr</td>
<td>Frontal cortex + Anterior cingulate + Precuneus cortex + Parietal cortex PIB mean intensities Whole Cerebellum reference PIB intensity</td>
</tr>
<tr>
<td></td>
<td>AV45 PET</td>
<td>ucerkeleyav45</td>
<td>Frontal + Anterior/Posterior cingulate + Lateral Temporal + Lateral Parietal weighted AV45 means 4</td>
</tr>
<tr>
<td></td>
<td>FDG PET</td>
<td>ucerkeleyfdg</td>
<td>Angular Gyrus + Left + Right Temporal Cortex + Posterior Cingular mean Bilateral FDG values 4</td>
</tr>
<tr>
<td>CSF</td>
<td>tau</td>
<td>upenmbergk3</td>
<td>tau concentration</td>
</tr>
<tr>
<td></td>
<td>p-tau</td>
<td></td>
<td>p-tau concentration</td>
</tr>
<tr>
<td></td>
<td>Aβ</td>
<td></td>
<td>Aβ42 concentration</td>
</tr>
</tbody>
</table>

| Genotype | APOE e4 | apoeers, apoe0g2 | Number of ApoE4 alleles the subject carries (0, 1 or 2) |

Table 3.1: Computing features from files variables. Each feature was computed from the equivalent values stored in a given file, similarly to what was employed in [5]. The name of the files presented are the names of the variables in the ADNIMERGE R package. It is worth noting that some features summarize a group of assessments, as is the case of CDR, which corresponds to the sum of 6 sub-scores of CDR test. MRI features comprises the volumes of specific regions of the brain. PET features consist on the mean intensity values across established regions. PIB specifically, consists on the standardized uptake value ratio, calculated by comparing values with the whole Cerebellum reference value. From clinical laboratory tests, Tau, P-Tau and Aβ42 concentration levels are directly extracted from the biomarkers file, which include the results from AlzBio3 immunoassay reagents on a Luminex 100 platform. The highlighted features are the features that are involved in the construction of the GRACE models.

These different types of data and other informations are presented to the user in different files. Each data entry in the files is identified with the respective roster id (RID) of the patient, the visit code (described in Table 3.3), the date of the particular exam and the phase cohort (ADNI 1, GO or 2) in which it is included.
The original dataset was obtained by merging all the files containing the necessary data (culminating in a R data frame exemplified in Fig. 3.3) which, without any pre-processing, comprises data from 2825 ADNI participants spread in 72600 data entries (described in Table 3.2).

![Figure 3.3: Random sample of the original dataset. Each entry corresponds to the value Y of a specific triplet of patient (RID), visit code (VISCODE) and assessment (Outcome). In the dataset we have information about the date in which the exam was performed (EXAMDATE) along with the register date (EXAMDATE.reg) and the cohort in which the data was assessed. Baseline information, not shown in this example, is also included and encompasses the age at baseline (AGE), the gender (PTGENDER) and the degree of education (PTEDUCAT). It is worth noting that if the Outcome is 'DXCHANGE', it corresponds to the clinical review visit and will be posteriorly used to assign a class to each instance.](image)

<table>
<thead>
<tr>
<th>RID</th>
<th>VISCODE</th>
<th>COLPROT</th>
<th>Outcome</th>
<th>EXAMDATE</th>
<th>EXAMDATE.reg</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>f</td>
<td>ADNI1</td>
<td>MMSE</td>
<td>19-08-2005</td>
<td>18-08-2005</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>sc</td>
<td>ADNI1</td>
<td>CDRSS</td>
<td>17-08-2005</td>
<td>17-08-2005</td>
<td>0</td>
</tr>
<tr>
<td>2240</td>
<td>m06</td>
<td>ADNI0G</td>
<td>WholeBrain</td>
<td>15-06-2011</td>
<td>15-06-2011</td>
<td>114.9432</td>
</tr>
<tr>
<td>2240</td>
<td>m06</td>
<td>ADNI0G</td>
<td>DXCHANGE</td>
<td>22-06-2011</td>
<td>15-06-2011</td>
<td>Stable MCI</td>
</tr>
<tr>
<td>2240</td>
<td>m12</td>
<td>ADNI2</td>
<td>ADAS13</td>
<td>01-03-2012</td>
<td>01-03-2012</td>
<td>14</td>
</tr>
<tr>
<td>5078</td>
<td>m03</td>
<td>ADNI2</td>
<td>Ventricles</td>
<td>28-05-2013</td>
<td>28-05-2013</td>
<td>30.6744</td>
</tr>
<tr>
<td>5078</td>
<td>m03</td>
<td>ADNI2</td>
<td>WholeBrain</td>
<td>28-05-2013</td>
<td>28-05-2013</td>
<td>123.5506</td>
</tr>
<tr>
<td>21-08-2013</td>
<td>ADAS13</td>
<td>21-08-2013</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.3: Random sample of the original dataset. Each entry corresponds to the value Y of a specific triplet of patient (RID), visit code (VISCODE) and assessment (Outcome). In the dataset we have information about the date in which the exam was performed (EXAMDATE) along with the register date (EXAMDATE.reg) and the cohort in which the data was assessed. Baseline information, not shown in this example, is also included and encompasses the age at baseline (AGE), the gender (PTGENDER) and the degree of education (PTEDUCAT). It is worth noting that if the Outcome is 'DXCHANGE', it corresponds to the clinical review visit and will be posteriorly used to assign a class to each instance.

<table>
<thead>
<tr>
<th>DX MV</th>
<th>NL</th>
<th>SMC</th>
<th>EMCI</th>
<th>LMCI</th>
<th>AD</th>
<th>DX Change MV</th>
<th>Exam Date MV</th>
<th>Visit Codes NC</th>
</tr>
</thead>
<tbody>
<tr>
<td>307</td>
<td>(0.4%)</td>
<td>20533</td>
<td>(28%)</td>
<td>2036</td>
<td>(3%)</td>
<td>12213</td>
<td>(17%)</td>
<td>27833</td>
</tr>
<tr>
<td>211</td>
<td>(7%)</td>
<td>612</td>
<td>(23%)</td>
<td>166</td>
<td>(6%)</td>
<td>459</td>
<td>(18%)</td>
<td>875</td>
</tr>
</tbody>
</table>

Table 3.2: Number of data entries and number of different subjects in each of the diagnosis categories (NL, SMC, EMCI, LMCI and AD). DX MV is the number of baseline diagnosis missing values. DX Change MV is the number of visits without a sporadic diagnostic review. Exam Date MV corresponds to the number of assessments without exam date specified. Visit Codes NC is the number of visit codes not considered. This include "sc","m03","uns1","scmri","nv" and phone review assessments (see Table 3.3). These values represent the number of subjects with at least one missing value in the considered variable. The number of subjects in which none of the assessments have a DXChange variable are 824, accounting for 1440 data entries. All subjects have at least one Visit code in their set of multiple assessments.

### 3.2.2 Instance Labelling Variable

The instance labelling variable gives a unique label to each assessment in order to then be used to aggregate assessments with the same label into instances. A label would ideally correspond to the exact date in which an assessment was performed, such that instances would correspond to a snapshot of a patient in a precise moment. However, this is unrealistic to the presented study since the scheduled exams cannot be performed in one single day, as it would overburden the patients, being often performed within a timespan of weeks. The exam schedule of the ADNI programs the approximated dates on which to perform each date-specific battery of examinations and gives a timespan of 2 weeks to perform all the examinations for that specific visit. However, some data features from a specific assessment were gathered over more than 2 weeks (1.22±2.65 weeks), probably due to patient's unavailability.

The exams performed in a scheduled time window are tagged with the same visit code (viscode) (Table 3.3), informing about the battery of tests on which that specific exam is included.
Table 3.3: Possible Visit Codes labels and respective description.

As such, we chose this code as the label for each instance. It is important to notice that data has a visit code already assigned to each entry, and therefore this could be used directly. However, the rule to assign an exam to an instance must be the exam date and upon data analysis, we discovered some anomalous cases (Fig. 3.4), showing a variability between exam dates with the same visit code and from the same patient that could not be ignored. To bypass those inconsistencies, we decided to continue using the visit code as the instance labelling variable but with some corrections specifically based on the exam date, by considering only visit codes limited to a given temporal window.

Figure 3.4: Example of an anomalous assigned visit code. Patient 51 has the first assessments in December 2005, in which was its baseline. The m06 assessments followed approximately 6 months after, in May 2006. However, it is observed some baseline assessments dating September 2011, a 6 year difference from the other assessments. Checking the registry of patient 51 it is possible to notice that what was perceived as baseline in the dataset was in fact the visit m72. The problem occurred from the transition to ADNI2.

At this point, we chose a temporal limit to decide whether an assessment is or not from a given visit. Short time windows lead to merging into instances only assessments with the same visit code and dates very close to each other, possibly dismissing more data, while wide time windows include a greater number of assessments with the same viscode but introduce error from exams with irregular dates.

The taken approach was based on the schedule defined in the ADNI protocol, in which visits are performed every 6 months, until reaching the 24th month, where it starts to be just annually visits with...
intermediary phone review assessments. To maximize the data quantity while still maintaining coherence with the ADNI study schedule, we decided to aggregate assessments limited to a maximum time difference of 6 months.

Particularly, the visit codes correction classifies each data entry depending on the time difference with the date associated to that visit (pivot date), extracted from the file Registry. Therefore, we attached a visit code regardless of the one already assigned to the data entries, relying on just the exam date and the information on Registry. For each data entry, the registry date closest to that entry’s exam date needs to be found, and if the difference between them is below 3 months, that entry is classified with the visit code of that pivot date (e.g. Fig. 3.5). If two exams of the same type of assessment are assigned to the same visit code, the exam which is closest to the pivot date is picked in detriment of the other.

![Figure 3.5](image)

**Figure 3.5:** Function of assigned visit code given an exam date for an exemplifying set of pivots dates. The registry file includes 5 visits, baseline, “m06”, “m12”, “m18” and “m24” and the respective registry dates which are denoted by the bullets under the time axis. As can be seen, the registry dates may not correspond to exactly 6, 12, 18 or 24 months after baseline. Here, for instance, it is observed that “m12” is slightly earlier than 12 months and “m18” and “m24” are slightly later. This leads to a shortening of the 3 months limit in the case of the “m12” and “m06”. The exam dates which have no pivot date from less than 3 months will be discarded as missing values (NA).

As it is depicted in Table 3.3, some visit codes cannot be included in the dataset since they do not fit in the formulations made so far. This include the viscodes of the screening visits, along with visits labeled as ‘no visit defined’ and phone review assessments (m30, m42, etc.).

Specifically, screening cannot be assumed as an independent visit code because it is not sufficiently separated from other visits. Screening is performed up to about a month before the baseline visit but can also be carried out in between other visits, due to change of ADNI phase or for verification. In the cases where it is not tagged as a screening fail, the screening gives extra information from a set of variables that can be used later in the classifier.

Consequently, we did not removed directly the unwanted viscodes from the preprocessing phase. Instead, we implemented in the viscode correction routine, the possibility to incorporate those assessment values, by removing registry pivot dates with those viscodes. In this way, the visits with less than 3 months of the assessments with unwanted viscodes are sought. Unwanted viscodes assessments are assigned to a approved viscode if in the 3 month range.
Regarding m03 visits, we opted to make the same approach as the screening visits. Visits on the 3rd month only occur in ADNI GO and 2 and correspond to a battery of purely imaging exams. Therefore, m03 instances would have every feature as a missing value except for the imaging set. Additionally, it is incoherent to have an instance representing a timespan of 6 months while others represent 3 months' windows. Therefore, m03 is ignored from the set of registry dates and the m03 exams will be assigned to a visit code, according to the 3 months threshold.

The resulting overall picture of corrected visit codes is depicted in Fig. 3.7. An example of corrected viscodes is represented in Fig. 3.6.

<table>
<thead>
<tr>
<th>RID</th>
<th>Outcome</th>
<th>EXAMDATE</th>
<th>EXAMDATE.reg</th>
<th>Y</th>
<th>VISCODE.pre</th>
<th>VISCODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>WholeBrain</td>
<td>23-07-2009</td>
<td>16-12-2008</td>
<td>104.6448</td>
<td>m36</td>
<td>NA</td>
</tr>
<tr>
<td>51</td>
<td>MMSE</td>
<td>30-05-2005</td>
<td>30-05-2006</td>
<td>28</td>
<td>m06</td>
<td>NA</td>
</tr>
<tr>
<td>51</td>
<td>Entorhinal</td>
<td>20-09-2011</td>
<td>08-12-2006</td>
<td>2655</td>
<td>bl</td>
<td>m72</td>
</tr>
<tr>
<td>51</td>
<td>Hippocampus</td>
<td>20-09-2011</td>
<td>08-12-2006</td>
<td>3.232091</td>
<td>bl</td>
<td>m72</td>
</tr>
<tr>
<td>51</td>
<td>Ventricles</td>
<td>20-09-2011</td>
<td>08-12-2006</td>
<td>30.78332</td>
<td>bl</td>
<td>m72</td>
</tr>
<tr>
<td>51</td>
<td>WholeBrain</td>
<td>20-09-2011</td>
<td>08-12-2005</td>
<td>103.0864</td>
<td>bl</td>
<td>m72</td>
</tr>
<tr>
<td>61</td>
<td>ABETA</td>
<td>05-08-2008</td>
<td>13-12-2007</td>
<td>172</td>
<td>m24</td>
<td>NA</td>
</tr>
<tr>
<td>61</td>
<td>PTAU</td>
<td>05-08-2008</td>
<td>13-12-2007</td>
<td>26</td>
<td>m24</td>
<td>NA</td>
</tr>
<tr>
<td>61</td>
<td>TAU</td>
<td>05-08-2008</td>
<td>13-12-2007</td>
<td>70</td>
<td>m24</td>
<td>NA</td>
</tr>
</tbody>
</table>

**Figure 3.6:** Example of 3 different visit code corrections. ‘VISCODE.pre’ is the viscode before the correction. Patient 51 (represented in Fig. 3.4) had some inconsistencies in bl visits, which were solved by assigning the closest visit code m72. The other two examples, subjects 47 and 61, had assessments overly separated from the register date. Here we can observe that m30 is ignored and some values, as is the case for the assessments of patient 61 performed in August 2008, will be dismissed.

### 3.2.3 Extracting Progression to AD

Another feature that was not incorporated in the GRACE algorithm was the longitudinal progress of the diagnosis status of a patient. The original study incorporated the diagnosis status at baseline, which was obtained through the file arm, but just employed it to compare and visualize patterns from different cohorts in the resulting curves.

However, in this work it is necessary to incorporate information about the diagnosis changes in order to identify the converters and non-converters MCI subjects. Similar to the visit codes, it was performed corrections to the diagnosis changes in order to avoid unnecessary data rejection and exam dates inconsistencies.
The diagnosis information is available in the file DXSum, in which the variable DXChange characterizes, when available, the current state of each patient at each visit. The outcome of the diagnosis review can be “Reversion: Dementia to MCI”, “Reversion: MCI to NL”, “Stable: NL”, “Conversion: NL to MCI”, “Stable: MCI”, “Conversion: NL to Dementia”, “Conversion: MCI to Dementia” or “Stable: Dementia”. It is important to notice that, contrary to the information expressed in the arm file, which includes the diagnosis at baseline, the variable DXChange from file DXSum does not differentiates SMC from other diagnosis categories. Instead, the patients with diagnosed SMC at baseline are considered “Stable: NL” and, in one case, “Conversion: NL to MCI”. Patients who reversed to a previous state are removed from the dataset, since they are considered to be misdiagnosed.

The DXChange values were assigned to each entry in a similar way of that applied to the visit codes, i.e. exams from which the dates are closest to the diagnostic clinical assessment date will be tagged with that diagnosis. We maintained the same 3 month limit to ensure that instances comprised only 6 month information. An example of corrected DXChange is represented in Fig. 3.8.
Figure 3.8: Correction of DXChange based on the comparison between the dataset exam date and DXSum exam date. As it can be observed, the viscode m18 differs in examdates from the Registry and DXSum files, therefore the DXChange obtained previously to the correction does not express the outcome of the closest diagnosis review, and thus is transformed. DXChange.pre represents the previous to correction DXChange and DXChange.cor is the corrected DXChange.

Additionally, missing values were filled when the complementary information to infer the value of the diagnosis change was available, as it is exemplified in Fig. 3.9. Namely, it was implemented a correction to fill DXChange missing values by assessing the diagnosis status at baseline. Some of the cases where the DXChange variable is missing are in fact baseline visits. Consequently, since diagnosis at baseline can be extracted from the file arm, this is used to infer the missing value of DXChange. In the file arm, diagnosis attains different outcome levels, including “NL”, “SMC”, “EMCI”, “LMCI” and “AD”. The transformation that I performed here was to convert baseline “NL” and “SMC” into “Stable: NL”, baseline “EMCI” and “LMCI” into “Stable: MCI” and “AD” into “Stable: Dementia”.

Figure 3.9: Correction of DXChange based on the diagnosis at baseline extracted from the file arm.

Based on the assumption of a sequence of diagnosis status in the progression of AD, i.e. from NC to MCI and then AD, the value for missing diagnostic data was inferred, such that assessments made between two diagnosis dates that reveal the same output will have that same output (Fig. 3.10). Similarly, assessments which are made prior to a diagnosed NL evaluation will be filled with “Stable: NL”, and for a diagnosis of stable dementia, the assessments made after and with missing diagnosis change value will be filled with “Stable: Dementia”.
Figure 3.10: Correction of DXChange based on the assumption of a continuous deterioration of diagnosis status. Assessments before a known assessment with diagnosed NL are “Stable: NL”, after a known assessment with diagnosed dementia are “Stable: Dementia” and between two assessments with the same diagnosis, the middle evaluations are that diagnosis status, as exemplified in the figure.

From the original dataset with 72600 data entries from 2825 distinct subjects, the presented corrections on data inconsistencies and missing values, yielded a dataset with 67315 data entries from 2150 subjects. The amount of correct and removed entries compared to the dataset resulting from no corrections is depicted in Fig. 3.11.

Figure 3.11: Diagram representing the number of data entries restored and removed when applying the proposed corrections. From the total number of 57517 data entries of the Dataset without correction, 56261 were considered in the corrected dataset, while 1256 were removed or altered. From that 1256, 775 data entries that were removed lead to an instance’s missing value, in contrary to 304 data entries which, by being removed, prevented the inclusion of 107 instances. The Dataset with Corrections comprises 67315 data entries from 2150 subjects, including 9147 from filled missing values, 1739 of filled mv that originated 50 new instances, 111 data entries with changed visit codes and 57 data entries with changed DXChange.

3.3 Computing Learning Examples

Once computed each assessment, in a single dataset, with the respective visit code and diagnosis status, we grouped data entries into classifiable instances, uniquely defined by the pair visit code/subject RID, and, with the diagnosis status, assigned each instance with the respective class: positive/MCI converters (MCInc) or negative/MCI non-converters (MCInc).
Finding if an instance corresponds to a conversion class involves comparing the instance with a later one, which serves as a reference point and from which is collected information about the moment of conversion.

The problem of choosing the reference instance can be approached from two perspectives. On one hand, it is possible to classify each instance according to the diagnostic differential between the first and last evaluations. However, this approach disregards the longitudinal changes that occur within the subject, i.e. early instances with a feature pattern significantly distinct from later instances are given to the classifier as belonging to the same class, which, in turn, introduces noise and inconsistencies to the training. Additionally, assessed patients with no conversion are perceived as non-converters, which cannot be precisely defined with the existing evidence.

On the other hand, an alternative approach is to designate a time period from which to consider a conversion (represented in Fig. 3.12), so that instances within a temporal window of the conversion date are considered converters, whereas instances that occurred out of the temporal window (N years before the conversion) are considered non-converters. This approach implies class assigning mediated by the disease stage of the given patient in a given moment (defined by a snapshot/instance), instead of the first and last approach, in which the class assigned was mediated by each patient quantity of available data. Here, the result is a dataset with different classes for instances of the same patient.

![Figure 3.12: Class labelling scheme. The variable threshold is represented by distance X. In this figure, X is considered a 3-years temporal window. The subjects for which there is no conversion assessed (top) may have the later instances within a temporal window of a potential future conversion. Therefore, those cases are labelled as “Unknown” due to insufficient information about conversion in the defined temporal window. Once detected a conversion date in the available assessments (bottom), data from less than X years before are tagged as “MCI-Converters”, otherwise being “MCI-Non-Converters”.

To decide which value to adopt as the threshold we considered two different goals. First, the temporal window must be chosen in a way that allows the training dataset to represent a balanced classes distributions. With this purpose, we plotted the distributions of classes per number of selected time thresholds
(Fig. 3.13), which shows that a time-window of approximately 2.5 years balances the data. The other point worth considering is the medical relevance of such temporal windows, which depends on the disease being study. AD is a long-term disease which can progress from more than 2 decades. Accordingly, prediction of AD within 1 or 2 years is insignificant, since by that time, physicians already have the discriminative power to attain a prognosis. According to medical background provided by expert colleagues specialized in AD clinical assessments, prediction within 3 years is significant. With this information and observing a sufficient balanced data, we opted to use a 3-years temporal window.

![Figure 3.13: Plot of the number of instances versus the number of years for the temporal window. The distribution of classes becomes balanced close to 2.5 years.](image)

### 3.4 GRACE Models

GRowth models by Alternating Conditional Expectation (GRACE) [89] were used to infer the longitudinal staging of the subject’s snapshots (instances). The algorithm consists on modelling monotone smooth curves of long-term progression of biomarker severity through time. Long-term data, however, is not available to directly fit a model curve, instead, data consists on longitudinal evaluations, reaching about 6 years, extracted from subjects in different disease stages. The study data comprises snapshots of biomarkers progression for each patient. When comparing snapshots within a subject, we can extract the short-term disease stage relatively to the baseline visit, but in a long-term time axis, a reference point is unknown and each patient cannot be anchored to a stage.

With the purpose of estimating the long-term growth curves along with inferring the unknown time-shift that each subject has experienced in the course of the disease, it is alternately performed the modelling of the data growth curve with a temporal shifting of each subject’s data according to the estimated curve. This shifting, in turn, reshapes the data such that a following estimation of a growth curve is updated. Consequently, from an initial short-term curve estimation, data is iteratively spread across a long-term time axis, to finish in a curve estimation of about 2 decades of AD progression.

The curve modelling is based on the assumption that the severity of any featured measurement behaves like an increasing curve. In fact, as long as the data used corresponds to subjects experiencing the AD pathological cascade, in early or later stages, it is conceptually valid to make that assumption, since
biomarkers are constantly deteriorating with the progression of the disease. Therefore, it is performed a smooth modelling of monotone functions for the biomarkers’ severity.

The implementation for modelling smooth growth curves, based on the work of Ramsay et al. [90], does not require pre-specifying the shape family of the curves. It is defined a class of monotone functions $f$, represented by the differential equation $D^2 f = w D f$, allowing the data fitting to be carried out for a nonparametric and even non-smooth function $w$.

The data model is defined by the linear Equation 3.1, which incorporates the value of the estimated curve function $g$ at time $t$, shifted by the subject-specific $\gamma_i$, along with the random-effects terms $\alpha$, a linear model of subject and biomarker-specific mixed-effects, and the residual error term, $\epsilon$. The model is a simplification of the classical invariant model (SIM), excluding the SIM rescaling parameters and including the additional random slope term $\alpha_{1ij}$:

$$Y_{ij}(t) = g_j(t + \gamma_i) + \alpha_{0ij} + \alpha_{1ij} t + \epsilon_{ij}(t) \quad (3.1)$$

Where $Y_{ij}(t)$ is the data point of subject $i$, from biomarker $j$, $t$ years after the subject’s baseline; $g_j$ is the progression curve specific for biomarker $j$; $\gamma_i$ the temporal shift specific for subject $i$; $\alpha_{0ij}$ and $\alpha_{1ij}$ are the parameters of the mixed-effect model for each pair subject/biomarker and $\epsilon_{ij}$ is the estimated error for each data point.

By alternatively estimating and updating the set of parameters, modelling is performed dividing the high dimensional problem into separated steps. Instead of trying to fit the data into the whole model, each parameter is fitted according to the data portion which is not being modelled by the other parameters, i.e. the partial residual of that parameter (Table 3.4).

<table>
<thead>
<tr>
<th>Partial Residual</th>
<th>Conditional expectation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{ij}^0(t)$</td>
<td>Long-term subject-specific smooth curve: $g_j(t + \gamma_i) = g_j(t + \gamma_i)$</td>
</tr>
<tr>
<td>$R_{ij}^1(t)$</td>
<td>Subject- and biomarker-specific intercept and slope: $g_j(t) = g_j(t) + \alpha_{0ij} + \alpha_{1ij} t$</td>
</tr>
<tr>
<td>$R_{ij}^2(t)$</td>
<td>Subject-specific time shift: $\gamma_i = g_j^{-1}(Y_{ij}(t)) - t$</td>
</tr>
</tbody>
</table>

**Table 3.4:** Partial residuals for each target parameter and the respective conditional expectations. The approximation of $E(R_{ij}^0(t) \mid g_j, t, \gamma_i)$ is made since here it is integrated over the function $g_j^{-1}$.

The two $\alpha$ parameters are modelled in a mixed-effect linear model incorporated in the iterative routine. So, estimation of the smooth growth curve is followed by mixed-effect linear modelling of the residual errors between values of data and of the estimated curves. This is performed in a cycle to reduce the amount of residual error (residual sum of squares). Then, the optimized set of biomarker-specific curves $g_j$ and mixed-effects are used to estimate the subject-specific time-shift $\gamma_i$. The time which better fits a given value $y$ in the constructed curves $g_j$, correspond to $g_j^{-1}(y)$. However, each biomarker $j$ gives its best fitting estimative of $t$. The resulting $\gamma_i$ corresponds to the mean of $\gamma_{ij}$ of the multiple biomarkers.
The algorithm outputs, for \( m \) subjects and \( n \) subjects, \( m \) smooth curves, \( m \times n \) sets of random effects estimates and \( n \) time-shifts.

**Algorithm 1: GRACE algorithm**

1. initialization: \( \gamma_i = 0 \)
2. while \( \text{RSS} = \sum_{ijt} [Y_{ij}(t) - g_j(t + \gamma_i) - \alpha_{0ij} - \alpha_{1ij}t]^2 > \text{tolerance} \) do
3. \hspace{.2in} initialization: \( \alpha_{0ij} = \alpha_{1ij} = 0 \)
4. \hspace{.2in} while \( \text{RSS}_j = \sum_{ijt} [Y_{ij}(t) - g_j(t + \gamma_i) - \alpha_{0ij} - \alpha_{1ij}t]^2 > \text{tolerance} \) do
5. \hspace{.4in} a. Estimate \( g_j \) by a monotone smooth of \( R_{gj}^j(t) \)
6. \hspace{.4in} b. Estimate \( \alpha_{0ij}, \alpha_{1ij} \) by the linear mixed model of \( R_{\alpha}^j(t) \)
7. \hspace{.2in} end
8. Set \( \alpha_{0ij} = \alpha_{1ij} = \epsilon_{ij}(t) = 0 \)
9. Given current set of \( g_j \), estimate each \( \gamma_i \) by averaging \( R_{\gamma}^j(t) \) over all \( j \) and \( t \)
10. end

**Figure 3.14:** Scheme of the first two GRACE iterations. In this example, values regarding the hippocampus volume are presented. Short-term data (A) is the initial dataset. From there a curve \( g_j \), along with the random-effects, are estimated (B). \( g_j^{-1} - t \) gives the value for the estimated \( \gamma_i \), which are applied to the dataset, shifting it from (A) to (C). The dataset in iteration 2 is updated and the iterative procedure continues until convergence.

The iterations are stopped either when the residual sum of squares is minimized, i.e. lower than the pre-specified tolerance (=10^{-3}), or when reaching to up to 10 iterations. This value was extracted from the previous work, where from simulated data it was concluded that 10 iterations reliably reconstructed the
original data [5]. Analysing the convergence of the GRACE iterative process (Appendix C - Fig. C.1), we observed that, for the model constructed from the whole training set, the process is stopped by reaching the tolerance level in the 8th iteration.

In order to estimate AD long-term curves, the data must represent a diverse population in different stages of the disease progression, i.e. the data set should incorporate various diagnosis categories, from normal to AD. However, since the aim of the estimated curves is to describe the pathology of the disease, subjects without any evidence of already being involved in the course of the disease, such as control subjects, are expected to introduce ambiguous information. In their paper, Donohue et al. [5] chose data from AD, LMCI and EMCI, along with normal controls who showed evidence of AD-pathology. Based on that previous selection, in this thesis, subjects with evidence of AD are defined by a lower accumulation of amyloid in the brain (Aβ < 192 pg/mL), PIB PET SUVR greater than 1.5, AV-45 PET SUVR greater than 1.1) and at least one ApoE ε4 allele. Furthermore, since in this work we include information about changes in diagnosis, it was possible to find the normal controls who converted to either MCI or AD, within the specified time window of 3 years. Therefore, the models will be constructed from data from 5 diagnosis/prognosis categories (Fig. 3.15), including 841 non-converters normal controls with evidence of AD (Prob.AD NLnc), 64 converters normal controls (NLc), 164 patients with MCI who did not convert (MCInc) and 219 who converted (MCIc), and 704 already demented patients (AD).

![Figure 3.15: Histogram of the different categories used in the construction of the GRACE models.](image)

3.4.1 GRACE Preprocessing

It is necessary to transform the ADNI data, in order to apply the GRACE algorithm. This was attained employing an already implemented routine to prepare data to be introduced in the GRACE algorithm. The format of the data (Fig. 3.16) to input into the algorithm requires the transformation of the following two variables:

1. Transformation of exam dates into centered “short-term” time $t$ represented by years from baseline;
2. Transformation of absolute values of the features into common scaled percentile values.

Regarding the former transformation, it is worth noting that the models are constructed according to within-subject changes in longitudinal features and assume between-subject interactions as completely
independent. In fact, the unknown relations between subjects ultimately are the latent variables we seek to model in order to extract estimated progression curves and subject-specific disease stages. Thus, as far as concerns the algorithm purpose, differences between participants’s age or exam date can be dismissed, since they do not infer any objective measurement of time (the patients are evaluated in different stages at the baseline, at different years). The time between assessments from a single patient though, cannot be modified, since there it lies an objective measurement of time of progression in AD pathology.

The temporal shift for each patient's data is removed, being the data of each patient centered to zero before the temporal shifting of GRACE. The time variable $t$ of each subject thus corresponds to the elapsed follow-up years since the baseline visit, centered in zero, i.e. for a subject with multiple follow-up visits, the baseline visit has $t < 0$, the last visit $t > 0$ and, if existing, the middle visit has $t = 0$. Different assessments cannot be centered independently though, since it would distort the timespan between assessments. Rather, MMSE dates are centered, and the other assessments are shifted accordingly, resulting in a $t$ corresponding to the difference, in years, between the date of an exam and the middle point between the baseline MMSE date and the date of the last MMSE assessment.

With this initialization, the algorithm starts with no presumption, reflecting the fact that the disease stage, in the “long-term” time scale, is empirically unknown and will be iteratively estimated through the course of the algorithm.

<table>
<thead>
<tr>
<th>id</th>
<th>viscode</th>
<th>group</th>
<th>argvals</th>
<th>outcome</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1377</td>
<td>m12</td>
<td>AD</td>
<td>-0.001368925</td>
<td>ADAS13</td>
<td>77.79822827</td>
</tr>
<tr>
<td>5012</td>
<td>m06</td>
<td>AD</td>
<td>-0.001368925</td>
<td>ADAS13</td>
<td>84.75001347</td>
</tr>
<tr>
<td>4382</td>
<td>m12</td>
<td>Prob AD NiNc</td>
<td>-0.001368925</td>
<td>ADAS13</td>
<td>9.138467788</td>
</tr>
<tr>
<td>1377</td>
<td>m12</td>
<td>AD</td>
<td>-0.001368925</td>
<td>CDRSB</td>
<td>77.76945652</td>
</tr>
</tbody>
</table>

**Figure 3.16:** Data sample in the required format to enter GRACE algorithm. Each entry correspond to a single triplet of patient (id), visit code (viscode) and biomarker (outcome). Each patient’s diagnosis category (group) is incorporated only to visualize the categories distribution over time after the estimation of the curves. The argvals consist on the model variable $t$. Remember that time zero represents the middle point of the participation timespan of each subject in MMSE follow-ups, so early dates, such as the baseline, will have negative $t$ values. The y values range from 0 to 100 according to a percentile transformation.

Regarding the data normalization into percentile scale, each assessment value is transformed to the respective percentile using the whole data, prior to the algorithm. Considering the unbalanced data distribution of the different diagnosis stages, percentiles were obtained by an empirical cumulative distribution (Fig. 3.17), weighted according to the inverse of the proportion of observations from each category. Notice that biomarkers expected to decrease with the disease, were adjusted in order to obtain an increase on biomarker severity with the decline of the measured value. This include MMSE, RAVLT, Hippocampus, Whole Brain and Entorhinal volumes, FDG and Aβ. The features expected to increase correspond to CDR-SB, ADAS-13, FAQ, Ventricles volume, PIB, AV-45, tau and p-tau. Notice that this scale transformation uses the whole dataset and therefore, we are modelling the GRACE features of the test set with a minimal dependency from the training instances. However, note that the weights employed for the percentile transformation represent the population prevalence of each category. If we were able to acquire the real prevalence, that would be the weight used, but, since we just have this dataset sample,
we estimated from it. Furthermore, note that the weights differ in diagnosis at baseline and thus, the bias derives only from EMCI and LMCI weights, since we only work with MCI instances.

![Figure 3.17: Percentile Transform Function: Cumulative Distribution across e.g. ADAS13 assessment results and Hippocampus ICV percentage. The vertical dashed lines correspond to the quartiles.](image)

### 3.4.2 Computing GRACE+ Learning Examples

The problem of classification implicates data partition into train/test subsets. Consequently, in order to avoid bias in the test set, GRACE models are constructed only from training data. After running the algorithm, each training data entry has the respective $\gamma$, resulting from the last iteration of the model. Additionally, for each data entry of subject $i$, from biomarker $j$ and time $t$, it is also available $g_j(t+\gamma_i)$, $\alpha_{0ij}$, $\alpha_{1ij}$ and $\varepsilon_{ij}(t)$. For the purpose of learning the classifier, we opted to not incorporate $g_j(t+\gamma_i)$ as a feature, since it is already depicted in the Baseline features, and neither the values of random effects parameters, since they do not demonstrate any obvious relation with conversion to AD. Disease stage, in other hand, is very appealing to introduce in AD prognosis’ data mining experiments. Besides being a very important measure to interpret in a clinical environment, disease stage, at least conceptually, correlates directly to years to onset of disease and thus years to conversion.

The parameters we appointed to be extracted from the GRACE model are the $\gamma_i$, representing the time-shift applied to the patient, and $t+\gamma_i$, representing the GRACE estimated disease stage. Instances from the same patient have the same $\gamma_i$ but differ in $t+\gamma_i$.

To assign test instances with the respective model parameters, $\gamma_i$ for each subject is extracted in a similar manner to that applied to training. However, for the test data entries, the time-shifts and disease stage are computed by calculating $\gamma$ partial residual using the long-term curves $g_j$ obtained in the model learned from training data, as represented in Fig. 3.18.
Figure 3.18: Processing of mapping $\gamma$ and $t+\gamma$ values for points in the test fold. In this example, values regarding the hippocampus volume are presented. As we can observe, the curve obtained in the last iteration is submitted to the test values, which, from the partial residual of that curve, will estimate all the other parameters, including the random effects $\alpha$ and the time-shift $\gamma$.

3.5 Classification Methodology

The objective of this thesis is to assess the improvements achieved when introducing GRACE parameters depicting the disease stage of a patient. In this context, the proposed learning method aims to find for each of the datasets, with and without the model features (GRACE+ and Baseline), the best predictor routines and then compare the resulting performance metrics.
Figure 3.19: Classification methodology data flow. The entire dataset is initially split (Fold Maker) into a set used for training with 75% of the entire dataset and an holdout test set with 25%. The training consists on a 10-fold CV grid-search routine with 75% of the total dataset. Cross-validation fold partition generates 10 non-overlapping folds and is performed using 5 different random seeds, which will lead to 5 repetitions for each dataset, i.e. more results and more robust conclusions. To assess GRACE+ results, a GRACE model is constructed from each of the 10 training folds, followed by merging the model variables with the training set (Train + γ) and submitting the curve function (g_j) to the test set, in order to compute the test with model variables (Test + γ). The 10-fold CV training is incorporated in an iterative cycle for the grid search within other cycle where it will be tested the different data Filter and SMOTE percentage combinations. The classifier/preprocessing that yields the best results in training is validated with the holdout set.

The employed data mining procedure (Fig. 3.19) does not follow a specific family of classifiers, parameters or preprocessing methods. Instead of pre-defining one single methodology, we test a set of different classifiers using distinctly preprocessed datasets. Training is performed in a 10-fold cross validation routine which uses a subset with 75% of the total dataset (detailed in Table 3.5). The pre-processing/classifier pair yielding the best 10-fold CV performance is chosen to represent each dataset and will be validated in the holdout set (25% of the total dataset, described in Table 3.5), from which we extract the results for further comparisons. To attain a more statistically powerful comparison, different cross-validation fold partition seeds were used to train the models. Stratified data partition was employed in order to maintain in each fold, the same class distribution as the whole dataset. To prevent a bias test set, the data partition was performed by inserting instances from the same subject in either the train or in the test set.

<table>
<thead>
<tr>
<th></th>
<th>MCInc (N=615)</th>
<th>MCInc (N=421)</th>
<th>Combined (N=1036)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>69.1/73.6/78.95</td>
<td>68.5/74/79.7</td>
<td>69.1/74/79.3</td>
</tr>
<tr>
<td>Gender: Male</td>
<td>58% (358)</td>
<td>67% (280)</td>
<td>62% (638)</td>
</tr>
<tr>
<td>Education</td>
<td>14/16/18</td>
<td>13/16/18</td>
<td>13.75/16/18</td>
</tr>
<tr>
<td>0</td>
<td>36% (221)</td>
<td>61% (255)</td>
<td>46% (476)</td>
</tr>
<tr>
<td>1</td>
<td>47% (291)</td>
<td>33% (141)</td>
<td>42% (432)</td>
</tr>
<tr>
<td>2</td>
<td>17% (103)</td>
<td>6% (25)</td>
<td>12% (128)</td>
</tr>
</tbody>
</table>

Table 3.5: Training set characteristics. This dataset corresponds to 75% of the whole dataset and it is from this dataset that the CV creates the 10 folds.
Table 3.6: Test set characteristics.

<table>
<thead>
<tr>
<th>Age</th>
<th>MCInc (N=210)</th>
<th>MCIrc (N=141)</th>
<th>Combined (N=351)</th>
</tr>
</thead>
<tbody>
<tr>
<td>70.9/76.4/80.4</td>
<td>66.5/73.9/79.2</td>
<td>70.5/75.3/79.3</td>
<td></td>
</tr>
<tr>
<td>Gender: Male</td>
<td>68% (143)</td>
<td>74% (104)</td>
<td>70% (247)</td>
</tr>
<tr>
<td>Education</td>
<td>16/16/18</td>
<td>14/16/18</td>
<td>15/16/18</td>
</tr>
<tr>
<td>ApoE4</td>
<td>0</td>
<td>30% (63)</td>
<td>65% (92)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>51% (108)</td>
<td>33% (47)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>19% (39)</td>
<td>1% (2)</td>
</tr>
</tbody>
</table>

Table 3.7: Parameters Grid Search.

The set of classifiers include some of the most common ones, i.e. C4.5 Decision Trees, K-Nearest Neighbour, Logistic Regression, Naïve Bayes, Random Forests and Support Vector Machines (with Gaussian and Polynomial Kernels) and are implemented in WEKA. The parameters for each classifier are optimized in a grid search (Table 3.7).

3.5.1 Classification Preprocessing Techniques

Different combinations from preprocessing filters are applied in order to detect and evaluate the variations in results when employing different data preparation methods. Namely, we tested 18 combinations, resulting from three different SMOTE percentages (0%, 50% and 75%), three distinct feature selection techniques and two ways of approaching missing values (with and without missing value imputation). The combination of preprocessing techniques, along with the classifier resulting in the best performance will be chosen to be validated. The best performance is ranked according to weighted AUC (discussed in Subsection 3.5.2).
less, feature selection is applied to analyse if the model parameters are between the resulting selected features. A prior exploratory analysis was performed and is shown in Appendix B.

FS in WEKA requires the predefinition of two aspects of the algorithm: the attribute evaluator, which is the way the algorithm will classify a value to each specific feature subset; and the attribute space search method, which allows to specify a methodology with lower computational burden than the default technique of searching all the possible combinations of features (exhaustive search). One standard way to bypass this exhaustive search approach is to use Best First search, which searches the feature subset space by greedy hillclimbing with a backtracking facility. If forward search is performed, the algorithm begins with a zero-feature set, iteratively starts adding the feature that if incorporated in the feature set yields the best results and stops when a pre-specified number of consecutive iterations showing no improvements has occurred. The search can also be backward, where the starting set is the complete feature set, or bi-directional, where the starting set is partially complete and evaluates the two directions, adding and removing features. Regarding the feature evaluator, there are two standard techniques: the wrapper method, which implies training a given classifier with a feature set and evaluate that feature set depending on the classifier's outcome; and the ranker method, which classifies each instance depending on a given metric, for instance the correlation with the dataset class attribute.

The FS used within the 10-fold CV routine was initially the wrapper attribute evaluator using the generalization performance of a Naïve Bayes classifier to evaluate a feature set, while searching the attribute space with the Best First approach. However, upon the analysis of results (Chapter 4), due to some poor choices made by this approach, we decided to use another completely distinct feature selection method, specifically, a correlation-based feature selection (Cfs) with a bi-directional Best First search approach, with 10 allowable consecutive non-improving nodes before terminating the search.

**Missing Values Imputation**

As we described in Fig. 3.20, the problem of handling missing values is critical for the target dataset. Nevertheless, it is important to notice that the existing missing values (MV) in this dataset are mostly concentrated in four or five features and therefore, with each classifier's underlying decisions and also when applying feature selection, the effect of MV is quite mitigated, since classifiers often opt to ignore those features and FS removes the uninformative, which include the highly incomplete ones.

The collection of CSF requires a Lumbar Puncture, consequently leading to CSF extraction being less recurrently scheduled. As for Aβ imaging procedures, PIB PET was discontinued from ADNI1 to ADNIGO and replaced by AV-45, which is only annually assessed and has been being collected just from visits in the last 6 years. Some missing values correspond to exams not included in a particular visit by the study protocol.

Regarding this dataset, the missing values are not significant for NM and sMRI features. In other hand, the great percentage of missing values found in CSF and Aβ imaging variables makes them unnecessary in the context of classification. This can be perceived when analysing the constant exclusion of this features after feature selection. However, in the context of GRACE curve modelling, it is relevant to include this type of measurements to incorporate information of the earliest phases of the disease, speculated to be the stage where this variables severity more rapidly increases.
Some classifier implementations have an internal way of dealing with missing values. SVMs and logistic regression use the median/mode imputation, decisions tree like C4.5 and Random Forest algorithms use statistical methods that minimize the missing data effects and kNN assumes the maximum possible distance in missing data cases. Nevertheless, missing value imputation (MVI) was applied since it gives us an insight on the effects of this technique as well as having the potential to improve some classifiers using the target dataset. MVI was applied with the ReplaceMissingValues filter implemented in WEKA, which replaces missing values for both numerical and nominal attributes by the means and modes from the existing values of that attribute in the training set. No differences are expected from SVM and logistic regression using this MVI.

![Figure 3.20: Missing values percentage for each variable.](image)

**Synthetic Minority Oversampling Technique (SMOTE)**

Although the instances show already an approximately balanced distribution (825 (59.5%) MCIc and 562 (40.5%) MCInc instances), the dataset can be even more balance when applying a SMOTE filter. We chose to experiment with SMOTE since the data is not completely balanced, expecting small improvements though. The tested percentages were 0% (no oversampling), 50% (resulting in a balanced dataset) and 75% (inversion of the imbalanced class), depicted in Fig. 3.21. The oversampled instances are inferred from 5 nearest neighbours. This balancing could also be achieved with other methods, including bootstrapping with repetitions by sampling instances weighted by the inverse of the respective class fraction.
3.5.2 Model Evaluation Metrics

There are several different ways to rate a classifier. The most intuitive is the precision accuracy, which corresponds to a fraction of the total number of instances from a test set that were correctly classified using a given classifier. This value, however, can easily be biased depending on the used dataset. For instance, if we use a simple classifier that classifies all the instances as positive and we have a dataset with a great percentage of positive instances, the prediction accuracy is high but is obvious that with a balanced dataset it would not be the case. Other metrics include the specificity and sensitivity, which are similar to prediction accuracy, but are not directly influenced by the class distribution of the dataset. Sensitivity corresponds to the portion of positive instances correctly classified as positive instances and specificity the portion of negative instances correctly classified as negative instances.

In this thesis, the performances are ranked according to the area under the receiver operating characteristic curve (AUC). This curve depicts the response of the classifier facing different decision thresholds. In other words, depending on the specific mechanisms underlying each classifier, we generally have a way to know the decision probability assigned by the classifier for an instance to be in the positive class. Knowing that probability we can order the instances by confidence of belonging to the positive class in a vector \( k \). Pre-defining a threshold means that all the instances with probability higher than the threshold are classified as positive, and negative otherwise. The intuitive threshold is 50%, i.e. if an instance has more probability to be in positive class than to be in the negative, it is classified as positive.

The ROC is constructed by plotting the sensitivity against 1-specificity. Each point in the ROC is computed by the pair (sensitivity, 1-specificity) of each possible threshold. The (0,0) correspond to considering no instance as positive and (1,1) corresponds to considering all instances as positive. If there are N positive instances and they fill the first N positions in \( k \), then it is a perfect ROC (AUC=1) and it means the classifier performed a perfect job at separating the two classes.

This is intuitively a good metric to use since it assesses the separability/overlap between the probability distributions of converters and non-converters. Increased AUC indicates that the two classes distributions are more significantly separated and thus the classifier is a better predictor across the features delimiting thresholds.
4.1 GRACE Progression Curves

In this section, we present and discuss the results from estimating the long-term AD progression curves with GRACE algorithm. The models presented here (shown in Fig. 4.1, along with the shifted data) are the result of a GRACE algorithm applied in the 75% training set. It is worth noticing that other 50 models were constructed, each one from a given train fold, with one of the 5 fold-maker random seeds. This results are presented in the Appendix C (Fig. C.6, Page C-6).

As already mentioned, different categories of diagnosis/prognosis were given to the algorithm, namely normal control subjects with evidence of AD pathology (by demonstrating abnormal values of Aβ, PIB, AV-45 or ApoE4) along with normal controls who converted within 3 years to MCI or AD, MCI converters and non-converters and AD patients. These categories are differentiated in Fig. 4.1 and we can observe a recurrent separation pattern from AD and normal controls. The number of data points from normal controls with some evidence of AD (Prob.AD NLnc) overcomes the other categories, leading to a great yellow area covering the plots. However, if observed more closely, it is detected a pattern of succession from the different categories. Values from Prob.AD NLnc define the earliest stages of the disease, followed by NL converters (NLC), MCI non-converters, MCI converters and culminating in AD patients, whose values occupy for almost all the markers the upper-right corner of the graph. The presence of this two extremes (NL and AD) successfully led to the reconstruction of an initial and a final plateau and thus the hypothesized logistic shape curves. Moreover, a slight difference between the plateau is observed, i.e. the early plateau is not as steep as the final plateau, which may indicate that the biomarkers slow down but actually do not undergo a stabilization in the latest stages of the disease and continue constantly deteriorating. This fact is very pronounced in the measures of brain activity, such as the volumes of the Ventricles and Hippocampus, and also FDG, which may indicate that neurodegeneration continues further into deep neuronal circuits, does not stop on functional and cognitive disability and can culminate in brain regions associated with more basic life sustaining functions. Besides appearing to contradict the hypothesized model of Jack et al. [4], it is important to notice that in fact the peak of rate of deterioration in brain neurodegeneration markers is found relatively at the same timepoint as the peak of neuropsychological tests deterioration rate. In fact, neurodegeneration curves show an almost linear increase, indicating that they start deteriorating at early stages and neuronal loss is persistently occurring at an almost constant rate. This corroborates with the proposed order of biomarker events. These results may indicate that, besides neurodegeneration beginning to be observable at approximately the same stage as neuropsychological performances starts to be notably worse, neurodegeneration progresses further, even after the cognitive/functional domain is already significantly impaired.
Figure 4.1: Progression Curves obtained from the GRACE model using the training dataset (75% of the whole dataset). Prob.AD NLnc=the normal controls with some evidence of AD, NLc=normal controls who convert later in the study, MCInc=MCI non-converters, MCIc=MCI converters, AD=demented patients.
Regarding the differences between neuropsychological biomarkers, the curves indicate that it is more likely to find an early stage subject with slightly reduced MMSE, RAVLT or ADAS scores comparing with CDRSB and FAQ. In fact, the latter show the more pronounced deterioration, progressing from the healthiest value observed in the whole population to the worse case observed. In contrast, MMSE, RAVLT and ADAS do not show such a steep curve, as subjects have higher probability of yielding an already imperfect cognitive ability in early stages, whether because of normal aging or influenced by a potential AD pathology. This resulting reductions in the percentile range of each curve are also explained by high between-subject variability, which tend to flatten the mean trajectory. This is quite noticeable in highly incomplete features, such as CSF proteomics and Amyloid imaging (PIB and AV-45), and also in the Whole Brain volume and RAVLT scores.

The highly sparse variables are unable to effectively contribute to the model construction and thus result in more flatten and unreliable curves. CSF biomarkers are expected to progress in a similar way as observed in other features curves, such as FAQ or Hippocampus, but shifted to an earlier stage. However, the peak of CSF deterioration rate, expected to occur before \( t=-5 \), i.e. before the observed peak increases in other features, which are roughly between \( t=-5 \) and \( t=5 \), was not depicted in the final curves.

In general, the algorithm produced curves with expected shapes, all very similar to the results obtained in [5]. It is worth noticing that, in comparison with the previous results, measures such as AV-45 and CSF biomarkers showed slight differences, which may be explained by the incorporation of values that were not being assessed previously, namely, normal controls who converted to MCI or AD. This intensification of early stages values led to a more pronounced early plateau in this measurements, which came to result in curves with a more apparent logistic shape. The A\( _\beta \) demonstrates an exception to these cases, since it progresses in a more quadratic shape rather than logistic and undergoes the steepest increase phase in later stages of the disease, which contradicts the expected from Jack's model [4], which states that A\( _\beta \) is the precursor of all other events that occur in AD pathology cascade. Once again, as we provide the model with an highly sparse set of values of A\( _\beta \), the conclusions to be extracted from this features are limited. Nevertheless, the shape of A\( _\beta \) goes accordingly to the obtained in the previous work, indicating a recurring problem from this dataset.

Contrary to the results published in [5], in this thesis we have not explored the efficiency of each marker in terms of the signal-to-noise ratio, performed by assessing rates of biomarker change standardized by residual standard deviation (SD). Nonetheless, it can provide with an important analysis to further investigate the nuances and aspects of these curves.

### 4.1.1 Disease Stage Separability

With the purpose of assessing the differentiating power of \( \gamma \) and \( t+\gamma \), the probabilities of correctly classifying each instance depending on the respective \( \gamma \) and \( t+\gamma \) were plotted across different thresholds (Fig. 4.2). Accordingly, instances with \( \gamma \) exceeding the threshold of 1.46 are classified as converters and non-converters otherwise. With this dichotomization according to a simple threshold, we can observe that, by itself, \( \gamma \) can be used as a quite reliable disease index to distinguish converters from non-converters, with an accuracy of approximately 85%. The \( t+\gamma \) yields a smaller prediction accuracy of 83%, but is still considerable higher than the ADAS, which represents the other features and yields 79.3% of accuracy. Note
that with no threshold the prediction rounds the 60% which correspond to the proportion of converters in
the whole dataset.

\[ \text{Class Distribution} \quad \text{Prediction Accuracy (\%)} \]

\[ \gamma \quad 84.9 \]

\[ \gamma' \quad 0.4 \]

\[ \text{Class Distribution} \quad \text{Prediction Accuracy (\%)} \]

\[ t + \gamma \quad 93.1 \]

\[ t + \gamma' \quad 0.29 \]

\[ \text{Class Distribution} \quad \text{Prediction Accuracy (\%)} \]

\[ \text{ADAS} \quad 93.3 \]

\[ \text{Class} \quad \text{MCc} \quad \text{MCinc} \]

\[ \text{MCc} \quad \text{MCinc} \]

Figure 4.2: Baseline Results across different classifiers with different SMOTE percentages.

4.2 Baseline vs GRACE+ Results

In this section, results from applying the proposed data mining approach over the two datasets (BL and GRACE+) are presented and discussed. Here, the aim is to compare the two approaches and assess differences in evaluation metrics. The metrics shown are the prediction accuracy, weighted AUC (AUCW), sensitivity and specificity. We consider AUCW to be the criteria to evaluate the predictive power of a classifier.

Two distinct results are presented: first, we will evaluate the training procedure by analysing the results obtained from the 10-fold CV routine, using 75% of the whole dataset (Section 4.2.1); secondly, the validation results, which emerged when testing the optimal trained classifiers in the holdout validation dataset, will be shown (Section 4.2.2).

4.2.1 10-fold CV Training Results

In Fig. 4.3, we illustrate the best 10-fold mean results for each classifier, separably depicting each dataset (BL and GRACE+) results. These results emerge from five 10-fold CV repetitions, using five different random seeds in the fold partition. The classifiers/preprocessing with no significant difference with the top classifier/preprocessing are shown in Appendix D (Fig. D.1 - BL and D.2 - GRACE+). The classifiers were learned using grid search to find the best set of parameters and the outcome is depicted in appendix' Table D.1 (Page D-3).
Figure 4.3: The 10-fold CV results using Baseline and GRACE+ dataset, across all the classifiers. The presented values are, for each classifier, the results of the preprocessing techniques combination which yielded the highest mean AUCW across all fold partition random seeds, using the Baseline (left) and the GRACE+ (right).

The most noticeable aspects of these results can be summarized by the following:

- Naive Bayes (NB), using a kernel estimator, is placed as the top classifier in both datasets, using no preprocessing technique. Nevertheless, it is closely followed by Logistic Regression (LOG) and Random Forest (RF);
- The effects of the preprocessing benefits the overall performance, but not for top classifiers, except BL Logistic;
- The weighted AUC (the target metric) does not yield considerable differences for any of the classifiers. In fact, the small AUCW differences detected do not occur in the top classifiers (NB, LOG, RF), which practically do not change;
- A significant difference is depicted in Sensitivity/Specificity. Namely, when using GRACE+, there is a considerable decrease in sensitivity, coupled with a considerable increase in specificity. This means more MCI non-converters are being correctly classified, while more MCI converters are being incorrectly classified;
- Regarding prediction accuracy, the results are pessimistic for the GRACE+ models, being quite noticeable a decrease in almost all the classifiers when using GRACE+;
- The small variance in AUCW obtained from employing different CV folds, using different random seeds, suggests that the proposed methodology generate quite robust results.
Table 4.1: 10-fold CV results using Baseline and GRACE+ dataset, across all the classifiers. The presented values are, for each classifier, the results of the preprocessing techniques combination which yielded the highest mean AUCW across all fold partition random seeds. Legend: m.GRACE - mean with GRACE+; m.BL - mean with BL; p-value - t-test p-value.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>AUCW</th>
<th>Prediction Accuracy</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m.Grace</td>
<td>m.BL</td>
<td>p-value</td>
<td>m.Grace</td>
</tr>
<tr>
<td>DT</td>
<td>0.828</td>
<td>0.841</td>
<td>0.164</td>
<td>75.426</td>
</tr>
<tr>
<td>kNN</td>
<td>0.876</td>
<td>0.864</td>
<td>0.030</td>
<td>80.744</td>
</tr>
<tr>
<td>LOG</td>
<td>0.906</td>
<td>0.907</td>
<td>0.559</td>
<td>81.400</td>
</tr>
<tr>
<td>NB</td>
<td>0.911</td>
<td>0.911</td>
<td>0.968</td>
<td>81.890</td>
</tr>
<tr>
<td>One R</td>
<td>0.756</td>
<td>0.758</td>
<td>0.692</td>
<td>73.450</td>
</tr>
<tr>
<td>RF</td>
<td>0.896</td>
<td>0.896</td>
<td>0.998</td>
<td>80.502</td>
</tr>
<tr>
<td>SVM G</td>
<td>0.834</td>
<td>0.821</td>
<td>0.012</td>
<td>82.622</td>
</tr>
<tr>
<td>SVM P</td>
<td>0.833</td>
<td>0.823</td>
<td>0.007</td>
<td>82.102</td>
</tr>
</tbody>
</table>

From the best classifier to the worse, with regards to AUCW, the order is the following: Naïve Bayes, Logistic Regression, Random Forest, kNN, Decision Tree, Polynomial SVM, Gaussian SVM and One Rule. However, note that many of the classifier/preprocessing combinations that were not selected as the best for a given classifier are in fact still better predictors than other classifiers (see Fig.D.1 and D.2).

NB as the best classifier is quite a positive result, since NB is a probabilistic classifier. Returning the classification probability, which will then infer the predicted class, at this type of medical problems, is crucial for the physician, enabling him to perceive the confidence attained by the classifier.

Overall, we see that was beneficial to incorporate the preprocessing optimization search, since for half the best trained classifiers, at least one preprocessing technique is employed. The effects of the preprocessing methods are further analysed in Section 4.3.

Regarding AUCW, it is disappointing to observe that no considerable difference was attained when using GRACE+. However, it is worth noticing (as can be found in Table 4.1) that no statistical significance is attained for the cases where AUCW is lower in GRACE+, contrasting with the 3 classifiers which benefited from GRACE+, yielding statistical significant increases in AUCW.

The major differences between the two datasets are observed in the specificity/sensitivity values. There is a tendency for the GRACE+ classifiers to yield higher specificity and lower sensitivity relatively to the BL classifiers. Here, two goals are pursued: first, we want small differences between sensitivity and specificity (|Sensitivity - Specificity|), so that the classifier is equally able to classify correctly MCI converters along with non-converters; second, we seek the highest values of sensitivity and specificity combined (Sensitivity + Specificity), which is a form of prediction accuracy not weighted by the bias of classes distributions (similar to F-measure). Contrasting with the prediction accuracy (acc), this alternative accuracy view does a better job at evaluating classifiers regardless of the bias given by the dataset. In the used dataset, this bias (60% converters vs 40%non-converters) is not significant though. These results are presented in Table 4.2.
Table 4.2: Differences of Specificity and Sensitivity ($\triangle$Specificity, $\triangle$Sensitivity) between GRACE+ and BL and differences between Sensitivity and Specificity within each dataset ($|\text{Sensitivity} - \text{Specificity}|$). Here the p-values are computed comparing the absolute values of the differences $\triangle$ as well as the absolute values of Sensitivity-Specificity.

| Classifier | $\triangle$Sensitivity (%) | $\triangle$Specificity (%) | p-value | $|\text{Sensitivity} - \text{Specificity}|$ (%) |
|------------|-----------------------------|-----------------------------|---------|---------------------------------------------|
| DT         | -13.2                       | +11.3                       | 0.293   | 23.5                                        |
| kNN        | 0.1                         | 4.2                         | 0.010   | 4.0                                         |
| LOG        | -2.3                        | 3.1                         | 0.281   | 4.4                                         |
| NB         | -7.0                        | 6.8                         | 0.784   | 6.5                                         |
| One R      | -22.7                       | 22.3                        | 0.740   | 22.5                                        |
| RF         | -10.7                       | 13.1                        | 0.053   | 9.4                                         |
| SVM G      | -1.4                        | 4.1                         | 0.022   | 8.6                                         |
| SVM P      | -6.1                        | 7.9                         | 0.044   | 12.4                                        |

NB undergoes a mirrored change of sensitivity (BL:86.2%,GRACE+:79.2%) relatively to the specificity (BL:79%,GRACE+:85.8%), i.e. the $|\text{Spe}-\text{Sen}|$ difference and the value of Sen+Spe are maintained. Not so direct, are the changes with Random Forest, from which is observed that $|\text{Spe}-\text{Se}|$ is significantly reduced using GRACE+ (BL:14.5%,GRACE+:9.4%,p-value=0.002) and that the gains in specificity ($\triangle\text{Spec}=+13.1\%$) make up (almost significantly, p-value=0.053) for the losses in sensitivity ($\triangle\text{Sens}=-10.7\%$). The Logistic Regression differences between Sensitivity and Specificity were not significant, but it was observed a larger disparity between those two metrics though. kNN was the classifier which in general, most benefited from incorporating GRACE features, having the specificity, prediction accuracy and AUCW significantly improved while yielding a small improvement in sensitivity.

After this specificity/sensitivity analysis, the relevance of the pessimistic prediction accuracy results is, to some extent, minimized. We observe that NB and RF have significant losses in prediction accuracy when using GRACE+, but we also observe that the specificity differences are quite close to the differences of sensitivity, if not higher. This means the classifiers have a better performance than the depicted by the prediction accuracy, i.e. GRACE+ may indeed reduce the prediction accuracy for this dataset, but does not yield worse results of specificity and sensitivity, indicating better generalization for different classes distributions.

In summary, we conclude that the results may not be as positive as expected but when closely interpreted, they reveal a robust evidence of considerable improvements in AUCW and specificity, which compensates the losses in prediction accuracy and sensitivity (see Table 4.2).

The evidence presented so far was the comparison of the best preprocessing combination for each classifier, and thus, a more profound analysis on the effects of GRACE+ can be made by incorporating all the preprocessing methods used to represent collectively each classifier. This analysis is shown in Fig. 4.4 and further complemented by Table 4.3.
The results follow a similar pattern as the results for the best classifiers (Fig. 4.3). Here, an increase on the size of the boxes is immediately observed, which results from incorporating the 5 different seed repetitions for each of the 18 combinations of preprocessing techniques. By including non-optimized classifiers into one single group, we are increasing within-classifier variance, which seems an unnecessary analysis, but it can provide us with a complementary big picture on how the classifiers reacted to the various preprocessing methodologies. The statistical power for comparing BL and GRACE+ is now greatly increased and therefore, statistical significant differences are altered when comparing with Table 4.1. Namely, we observe that when considering all preprocessing techniques, the GRACE+ classifiers appear to be worse than previously perceived, which goes against what was expected. Regarding the preprocessing techniques, missing values imputation is conceptually expected to attain no significantly different effect between the two datasets, since the only difference between the two is the incorporation of two features with no missing values. Therefore, the subsequent differences from preprocessing lay in the other two preprocessing techniques. We can observe in Fig. 4.3 though, that the optimized preprocessing for the GRACE+ classifiers for most cases is quite similar to the one used for BL classifiers. We see that BL kNN employs all techniques, but GRACE+ kNN does not employ SMOTE; as is the case for Gaussian SVM which uses feature selection with BL but not with GRACE+; and also for Logistic Regression,
which uses all the techniques with BL but none of them for GRACE+. This suggest that some underlying process from this two techniques is interacting with the GRACE features in the learning process. This is further investigated in more detail in Section 4.3.

4.2.2 Holdout Test Validation Results

In this subsection, we present the validation of the best trained classifiers discussed in the last subsection, as well as a big picture on the classifiers’ validation using all of the proposed preprocessing techniques, similarly to the previous subsection. The best classifier, as referred before, is the Naïve Bayes without preprocessing, for both GRACE+ and BL. The validation results of this classifier are represented in Fig. 4.5. Note that for validation, the results did not emerge from multiple repetitions, in contrary to training from which we had, for each experiment, five repetitions results. In this case, the results arrive from testing the classifiers in the holdout 25% validation dataset. The validation test is expected to result in overall worse scores than the training, since this dataset is much more unbiased and did not directly influence the training in any aspect. Thus, the classifiers are tested as whether they have small decreases and generalize well or great decreases and are poorer generalizers using new datasets.

Regarding the effects of GRACE+, we can observe that the validation results are relatively more positive than the training values, i.e. the overall results show that the expected decrease in validation is mitigated when using GRACE+. Relatively to BL, GRACE+ prediction accuracy undergoes a quite insignificant decrease (82.05% to 81.77%), while AUCW yields a relatively greater value (88.66% to 89.98%). Sensitivity, consistently with the training, decreases with GRACE+ (82.85% to 79.95%) but is compensated by a larger increase in specificity (80.85% to 85.82%).

![Figure 4.5: Validation results using Baseline and GRACE+ datasets for the best trained classifier, Naïve Bayes without preprocessing.](image)

Validation results indicate that the NB classifier is a robust technique, as the values with this holdout dataset lie close to the values obtained from training (Fig. 4.3). Furthermore, comparing the two types of results, we observe that GRACE+ undergo a smaller decrease when generalizing. For instance, with GRACE+, specificity maintained the same for both train (85.8%) and validation (85.8%), sensitivity just slightly changed (from 79.2% to 79.04%), as well as prediction accuracy (from 81.89% to 81.77%) and
AUCW (from 91.1% to 89.98%). Regarding the BL dataset, such generalization power is not observed, as BL dataset yielded quite better results in the training relatively to the validation: prediction accuracy went from 83.28% in training to 82.05% in validation, AUCW from 91.1% to 88.7%, sensitivity from 86.2% to 82.86%. Specificity, surprisingly, increases in the validation test, going from 79% to 80.9%, indicating that the BL models were already, to some extend, quite good generalizers.

![Figure 4.6: Validation results using BL and GRACE+ datasets, across the top classifiers which yielded the highest AUCW in training.](image)

**Table 4.4:** Differences between the training results and the validation results. △ represents the difference of those values from GRACE+ and BL. Positive △ means GRACE+ generalized better than BL, i.e. relatively to BL, the classifier attained a lower reduction from training to validation. (*) classifier which yields better results in training using BL, and in validation using GRACE+. (**) classifier which yields better results in training using GRACE+, and in validation using BL.

<table>
<thead>
<tr>
<th>Metric</th>
<th>BL</th>
<th>GRACE+</th>
<th>∆</th>
<th>BL</th>
<th>GRACE+</th>
<th>∆</th>
<th>BL</th>
<th>GRACE+</th>
<th>∆</th>
<th>BL</th>
<th>GRACE+</th>
<th>∆</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUCW (%)</td>
<td>-33.89</td>
<td>-20.18</td>
<td>13.71*</td>
<td>-2.18</td>
<td>-1.99</td>
<td>0.19</td>
<td>-2.48</td>
<td>-2.73</td>
<td>0.25</td>
<td>-4.6</td>
<td>-1.6</td>
<td>3.0</td>
</tr>
<tr>
<td>Prediction</td>
<td>-30.87</td>
<td>-17.6</td>
<td>13.27*</td>
<td>-1.2</td>
<td>1.03</td>
<td>2.22</td>
<td>-3.71</td>
<td>-0.12</td>
<td>3.63</td>
<td>-4.44</td>
<td>1.11</td>
<td>3.31</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>-15.74</td>
<td>-2.2</td>
<td>13.54</td>
<td>3.14</td>
<td>1.75</td>
<td>2.22</td>
<td>0.66</td>
<td>0.09</td>
<td>4.33</td>
<td>1.11</td>
<td>2.11</td>
<td>0.07</td>
</tr>
<tr>
<td>Specificity</td>
<td>-41.63</td>
<td>-27.6</td>
<td>13.24</td>
<td>-4.19</td>
<td>0.5</td>
<td>4.69</td>
<td>-6.66</td>
<td>-0.20</td>
<td>3.97*</td>
<td>-1.79</td>
<td>-8.07</td>
<td>0.22</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>71.37*</td>
<td>67.63</td>
<td>3.74</td>
<td>4.69</td>
<td>9.42*</td>
<td>3.16</td>
<td>12.74</td>
<td>-6.67</td>
<td>8.06</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The validation results from the different classifiers are shown in Fig. 4.6. With this illustration, along with the comparisons between training and validation results presented in Table 4.4, we can observe quite good results with GRACE+ comparing to BL. A majority of classifiers yield better generalization power with GRACE+ than using BL. Small differences are observed between train and validation for most of the classifiers, except DT and SVM's, which indicates that most classifier did not overfit to the dataset.
The different aspects can be summarized in the following:

- An overall improved result using GRACE+ is observed when comparing to BL. This is not the case for prediction accuracy of RF and sensitivity of NB and RF.

- Sensitivity, which was recurrantly worse with GRACE+, demonstrates a different pattern for some classifiers, including kNN and Logistic Regression, in which Grace+ attains greater values than BL;

- Decision Tree and SVM’s perform a poor generalization, as we observe a significant decrease from training to validation results, in both BL and GRACE+ (as can be seen in Table 4.4). For instance, DT, which yielded AUCW of 82.8% (GRACE) and 84.1% (BL) in training, culminates in 62.6% and 50.2%, respectively, with the validation dataset;

- kNN, Random Forest and Naïve Bayes perform well in generalizing the classification and this ability is not significantly altered using GRACE+ instead of BL;

- Logistic regression proves to be an exceptionally good generalizer when using GRACE+. Logistic regression AUCW is actually greater in the validation test (91.47%) than in the training (90.57%);

- Specificity manifests the same pattern as the training set, being higher in all classifiers when using GRACE+ instead of BL.

To assess results variability from the different preprocessing techniques, we show the deviation of the classifiers’ performance including all the preprocessing combinations, and thus we are able to analyse the effect these techniques have in the variance of the validation results (Fig. 4.7).

![Figure 4.7: Validation results using Baseline and GRACE+ datasets, across all the classifiers, including all preprocessing techniques combination for each classifier. Some of the poorer results are not shown in order to adjust the scale of visualization.](image-url)
As we can see, from Fig. 4.7 and complementary Table 4.5, the validation values indicate great generalization for the GRACE+ models, as well as an increased robustness from different preprocessing methodologies. This indicates that, for a general dataset, estimating each patient's disease stage and incorporate it in a MCI prognosis data mining approach is favourable for the purpose of attaining improved predictors. In fact, this greater likelihood of the model to attain good results in a general dataset is emphasized by two aspects of this analysis: first, because GRACE+ attains recurrently better results in a validation dataset and second, because GRACE+ exhibits smaller variance around different preprocessing techniques, whose benefits are known to depend on the given dataset.

### 4.3 Preprocessing Results

In this section, the outcomes of the classifiers using different preprocessing techniques are presented in more detail. Namely, the effects of SMOTE, missing values imputation and feature selection on the results will be assessed individually.

#### 4.3.1 SMOTE Percentage

In order to assess the potential improvements resulting from training with different classes distribution, a small set of synthetic oversampling percentages (0%, 50% and 75%) was tested, resulting in using the: 1) original dataset, 2) balanced dataset and 3) dataset with inverted classes distribution, respectively (see Fig. 3.21).

As expected, since the original dataset was already quite balanced (~60% MCIc vs 40% MCInc), oversampling to further balance the data did not yield significant improvements in the results, as can be observed in Fig. 4.8.
Figure 4.8: Results across multiple classifiers with different SMOTE percentages and the other techniques as the optimized in training.

When analysing the classifiers which benefited from applying SMOTE (decision tree, kNN, logistic regression and SVM), we observed that in fact, the differences for most are not that noticeable, except for the decision trees, which yield a more pronounced improvement in AUCW when using SMOTE. Statistical significant improvements comparing to 0% SMOTE are observed when using 50% and 75% with GRACE+ and using 50% with BL. Logistic regression using GRACE+ yields significant decreases in AUCW when applying both 50% and 75% SMOTE comparing to 0%. Regarding the best classifier (NB), using the 2 SMOTE percentages does not yield significant improvements, being the three NB classifiers in the top 3, for both GRACE+ and BL.

4.3.2 Missing Values Imputation

Regarding missing values imputation (MVI), the resulting improvements were according to expected, i.e. were not significant for most of the classifiers. The missing values percentages are either not signif-
icant (0-25%) or considerably high (80-95%) and thus, missing value imputation results in features with 80-95% of the instances yielding the same value, the mean of the 5-20% of non-missing values.

Furthermore, some classifiers have internal ways to deal with missing values, including the SVM’s and logistic regression which replaces missing values with the mean/mode, which are similar to the missing values imputation filter used. Therefore these classifiers culminate in the same results with and without this preprocessing technique.

NB and Decision Trees yielded a significant AUCW decrease when applying missing values imputation, for both GRACE+ and BL. NB assesses marginal probabilities for each feature and hence, missing values do not present a major obstacle since the computations are performed with the proportion of a given value to occur within the given non-missing value set. Therefore, filling missing values is clearly adding noise to the classifier, since the real values (non-missing) have now a lesser contribution to the probability due to an increase on the overall number of values, while the generated values have no sig-

Figure 4.9: Results across multiple classifiers with and without missing values imputation (mvi) and the other techniques as the optimized in training.
significant contribution since they are all the same value. For C4.5 Decision Trees implemented in WEKA, instances with a given missing attribute reaching that attribute node are split into fractional instances and placed in each leaf according to the frequencies of the observed non-missing values. Similar to NB, in DT, missing values are regarded according to the distributions of the non-missing values. Therefore, when filling the missing-values with mvi, the values are no longer split, but rather included into one leaf, carrying with it an amount of unwanted noise, which was mitigated when introducing the fractional instances. In Random Forest, over-fitting effects are mitigated and thus, the influence of mvi is not so notable, resulting in a lower and not significant decrease.

Contrasting with these two classifiers, kNN showed a statistically significant improved performance when applying mvi, which goes according to the expect, since many close neighbours were not being considered due to not being complete, with the original percentage of missing values. With mvi though, these close informative neighbours appear to be relevant for inferring/estimating the class.

### 4.3.3 Feature Selection

Feature selection was performed using two different techniques, including a wrapper evaluator according to subset performance in a NB classifier and a correlation-based feature selection. The initial aim was to perform a single simple feature selection to perceive the frequency in which features were being selected and whether the results could be maintained or improved. However, the employed single approach yielded unsatisfactory results. Therefore, it was performed another feature selection using the correlation-based and the results were significantly improved.

![Feature selection frequency applying the two employed FS techniques. The frequency corresponds to the number of times each feature was selected from the total number of repetitions (300) - for the 3 SMOTE percentages, for mvi and no mvi, for each of the 5 seeds and each of the 10 folds (3×2×5×10).](image)

**Figure 4.10:** Feature selection frequency applying the two employed FS techniques. The frequency corresponds to the number of times each feature was selected from the total number of repetitions (300) - for the 3 SMOTE percentages, for mvi and no mvi, for each of the 5 seeds and each of the 10 folds (3×2×5×10).
In Fig. 4.10, the frequency each feature is selected using both techniques is depicted. As it can be observed, adding to the results presented in Fig. B.2, which show that the neuropsychological features were amongst the more predictive features, the features that were more times picked from both techniques in BL were ADAS-13 (selected 600 times), ApoE4 (582 times), FDG (528 times), CDRSB (522 times), MMSE (450 times) and FAQ (390 times). The least times selected features were PIB (72 times), p-tau (123 times) and gender (138 times).

When the GRACE features are included in the learning process, an immediate aspect observed is the shift of feature selection towards the model parameters to the detriment of other important AD-related variables, including Hippocampus volume (selected 0 times), MMSE (144 times), CDRSB (108 times) and RAVLT (0 times). It is important to notice that $\gamma$ is not estimated from age, gender, education and ApoE4 number of alleles. In view of this, is natural for the feature selection to detect less redundancy between these features and the model features, and thus pick both, in detriment of the features used to construct the GRACE curves.

The $\gamma$ was selected in 100% of the repetitions and $t+\gamma$ was selected in 50% (just by Cfs). This may be explained by the way the attribute set is searched. Best first probably included $\gamma$ immediately in the first iteration and for the following iterations, adding $t+\gamma$ may not have been as useful for NB, since it was already learning with a quite similar feature. In this context though, Cfs should also have behaved similarly, since not only the feature set evaluation is performed by correlation with prognosis class, but features are also assessed by the redundancy they insert by being incorporated in a given dataset.

As already mentioned, feature selection using the NB wrapper evaluator was quite disappointing and led to a second experiment using another feature selection method. This may be explained when analysing the features that were recurrently selected by NB wrapper and not by Cfs, which included gender, age, Ventricles, Whole Brain and highly incomplete features, such as A$\beta$, p-tau, tau and PIB. In comparison with other features, such as the Hippocampus, MMSE, RAVLT or FAQ, the features selected with NB wrapper are conceptually poorer predictors, which may explain the poorer results when classifying instances.

From Fig. 4.11, we can observe statistically significant improvements in AUCW when employing Cfs in kNN and DT. Comparing Cfs results with no fs results, kNN gained 1.8% with BL and 0.7% with GRACE+, and DT increased in 2.17% with BL and 4.4% with GRACE+. As for the remaining classifiers, significant decreases were observed for SVM’s, Random Forest, which decrease in 1.5% and 2.4% AUCW when using Cfs, comparing to no fs, for BL and GRACE+ respectively, and Naïve Bayes, which shows a significant decrease of 1.6% AUCW for BL. It is important to notice though, that in contrary to this classifiers, logistic regression AUCW did not decrease significantly (p=0.27), indicating that this type of classifier can be simplified and to some extend, stay equally predictive.
4.4 Summary

The results presented in this chapter, while providing a descriptive analysis of the used ADNI dataset, describe the outcomes of the proposed approaches when predicting the progression from MCI to AD. Here, we present the summarized achievements with our approach.

This study was performed with the purpose of assessing the power of an estimated disease stage to separate MCI who converted to AD from those who did not, and also investigate the benefits yielded by incorporating that type of feature into a predictive classifier. Along with these findings, since the estimation of disease stage is performed by fitting long-term progression curves of different AD-key biomarkers, an illustration of the way in which AD pathological cascade becomes evident under different forms of biomarkers and the order in which these biomarkers become abnormal was also an important achievement and a means of validating previously published work [5].
The progression curves show an overall agreement with the results of previous work. From the observation of this curves (Fig. 4.1, Page 59), we were able to draw the following conclusions:

1. A noticeable separation between diagnosis/prognosis categories is observed, namely, according to the expected sequence pattern. NL non-converters are placed in the earliest phase of the disease, followed by the NL converters. MCI non-converters and converters are quite differentiated in the long-term timeline, being the latter placed very close to the latest stage which corresponds to AD.

2. Some curves cover the complete range of values, while others are significantly flattened. From NM features, FAQ and CDR-SB curves successfully describe the whole value range, whereas MMSE, ADAS and RAVLT do not. This indicates that is more likely to find a cognitively normal subject with imperfect results in the latter features and that the former features can be expected to more reliably distinguish subjects in early stages of the disease.

3. Elevated percentages of missing values lead to flatten progression curves, i.e. the estimated curves are not able to model the whole range of values for highly sparse features, since these features have high between-subject variability and are quite uninformative comparing to the others. These features include amyloid-imaging values (PIB and AV-45) and CSF proteomic levels (Aβ, tau and phosphorylated tau).

4. Neurodegeneration, measured by FDG-PET and MRI volume, closely tracks neuropsychological symptoms, assessed with NM, such as ADAS or MMSE, and functional impairment, evaluated with FAQ. Neurodegeneration curves though, by behaving in an almost linear fashion, suggest that neuronal loss starts before the onset of neuropsychological symptoms.

5. The shape of the curves in the final stages indicates that neurodegeneration does not slow-down into a plateau like cognitive function do, suggesting that the brain areas associated with memory, cognitive and functional abilities are just the initial areas targeted by AD, and that another brain areas will experience neuronal loss in future phases.

6. From CSF curves, no robust conclusion can be drawn. However, we see a logistic shape curve for most of the features with long plateaus, indicating that the deterioration phase is confined to a short phase, where the value go from normal to abnormal.

7. The Aβ curves indicate a completely contrary perspective from the proposed hypothesized model, defended by Jack et al. [4]. Here, the peak of growth occurs in the last phases of the disease, which, in fact, was expected to be the earliest event on the AD pathological cascade.

When using the GRACE extracted estimates of the disease stage in the proposed classification approach, the results, described in detail in the last sections, are overall pointing to the following implications:

1. The baseline dataset yields good results: the best trained classifier (Naïve Bayes without pre-processing) achieved 91.1% of weighted AUC, 83.28% of prediction accuracy, 79% of specificity and 86.2% of sensitivity. In particular, we observe that the default set already includes the key factors of AD and summarizes lots of information from different more specific values of NM or imaging data. Therefore, good results were already expected for the baseline.

2. The best classifier in training with the disease stage incorporated yielded 91.1% AUCW, 81.89%
prediction accuracy, 85.8% specificity and 79.2% sensitivity.

3. For the different classifiers, GRACE+ results in training (Fig. 4.3, Page 62) did not yield considerable improvements. AUCW for most cases has decreased, but not significantly, whereas the improvements observed are statistically significant. Across all classifiers, prediction accuracy and sensitivity underwent an overall significant decrease, whereas specificity was significantly improved.

4. The validation outcome using the best trained classifier (Fig. 4.5, Page 66) are more positive than the trained results, indicating a better generalization when using GRACE+.
   - AUCW: GRACE+ yielded 89.98% and BL 88.66%;
   - Prediction accuracy: GRACE+ yielded 81.77% and BL 82.05%;
   - Sensitivity: GRACE+ yielded 82.85% and BL 79.95%;
   - Specificity: GRACE+ yielded 85.82% and BL 80.85%;

5. In the big picture of the whole set of classifiers, the improvements using GRACE+ in validation (Fig. 4.6, Page 67) are quite more noticeable against the training outcome. Namely, we do not observe a pronounced decrease in sensitivity neither in prediction accuracy as was detected in the training results. In fact, these metrics are often increased with GRACE+, along with AUCW and specificity, which were already improved with GRACE+ in train and in validation are consistently improved.

This set of results provides initial empirical evidence of the benefits of incorporating the disease stage in a data mining approach to detect conversions from MCI to AD. Specifically, this study suggests that it is useful to frame the dataset in a long-term disease timeline constructed with the GRACE algorithm and from there, predict conversion.
Conclusion and Future Work
There is an increasingly necessity to unveil the underlying sequence of biological events occurring within the AD pathological cascade. This urge emerges from the fact that AD remains incurable, while facing a large increase of prevalence, due to the increase in life expectancy, decline in physical activities and unbalanced eating habits. The known risk factors include age, dietary options, lifestyle and genetic factors, but the exact mechanisms triggered by this factors are still unknown. The discovery of the sequence of events that precede the devastating neuropsychological symptoms caused by AD, along with accurate prediction of the subjects who are going to progress to dementia, is of major importance in order to efficiently carry out clinical trials and further research. However, due to limited investment in research, the existing data from longitudinal studies does not follow patients through the entire course of the disease, covering just a small portion of that period, depending on the moment the subject started the study. The investment on this type of research, however, is gradually expanding, increasing the number of AD studies and available data. In response to this escalation, an initiative was projected to develop standardized data collection methods and support data sharing, the AD Neuroimaging Initiative (ADNI). The data from ADNI includes neuropsychological, genetics, imaging and biofluid data, along with a large spectrum of other complementary information, and thus provide researchers with a great opportunity to test hypothesis and develop models of AD. In fact, ADNI data is being widely used worldwide and is an important milestone to future work in AD. In this context, this thesis was propelled with the aim of obtaining a big picture of the current research topics being approached with the ADNI data. The accomplishment of a general ADNI literature review led to the conclusion that two main topics were the focus of investigation:

1. Prediction of AD: a great amount of studies is focusing on using biomarkers data to predict the prognosis of a patient and to classify each patient with a diagnosis category. It is of major importance for clinical trials, to choose patients at early stages of the disease, and therefore, a tool to help decide with confidence if a certain patient is or not in an early stage of AD is crucial;
2. Longitudinal models: a big current research trend focus on collecting and analysing evidence to support hypotheses on the progression of biomarkers and the order of that sequence. This constitute a decisive field of investigation in order to further understand the underlying mechanisms of AD and bring out new ideas and hypothesis regarding the biological aspects and the connections with the known symptoms.

In this context, we used the ADNI dataset and decided to incorporate in one single methodology this two main research fronts, by tackling the problem of prediction of conversion from MCI to AD within 3 years, using features that were extracted from the construction of longitudinal models, namely, the estimated disease stage. Therefore this work had two main challenges: 1) to investigate the sequence in which biomarkers become abnormal and the longitudinal shape of that progression; 2) to assess the benefits of incorporating long-term disease stage feature, in a classification approach for prediction of conversion of MCI to AD.

For the first objective, the studies on longitudinal models have been found point in one of two major directions: 1) determining the sequence of biomarker events [69, 71] and 2) modelling the longitudinal progression biomarker curves [5]. Both perspectives can yield a disease stage for each patient, the former a discrete feature, and the later in a continuous timeline. Since the later approach provides a
more complete illustration on the way biomarker progress over time, we opted to use it by recurring to the GRACE algorithm. GRACE (or Growth curves by Alternating Conditional Expectation) was proposed by Donohue et al. [5] and is an algorithm aiming to fit growth curves of the biomarkers data values through time, while adjusting the time variable using the estimated curves. This iterative dataset transformation shifts each patient’s data accordingly to that patient’s performance across the curves of the entire set of biomarkers and culminates in an iteration where no further shifting is necessary to reduce the error on the curve estimations.

In this work, we were successfully able to reproduce the curves obtained from the previous work of Donohue et al. [5], with GRACE, using a different ADNI dataset. This was accomplished by obtaining and understanding the implementation in R of the ADNIMERGE and GRACE packages, and complementing it with a new data preprocessing and correction routines, and some adjustments to incorporate the results into the classification methodology. The dataset had to be augmented with two new variables, namely, an instance labelling variable and the registered progression to AD. The necessity for the implementation of a new preprocessing emerged from analysing the dataset and observing some inconsistencies, specifically in those two variables, which we were able to rectify.

This work resulted in a set of curves that goes accordingly to the previous work, which was expected, since we are using a quite overlapping dataset, and thus can be used to some extend as a validation of that work. Nevertheless, the fact that we were able to obtain similar results to previous work when including certain adjustments and adapting the algorithm for the particular problem of this thesis points to the fact that we are successfully testing that algorithm in our approach. As future work, one could design a GRACE methodology using different biomarkers, such as fMRI or DTI, in order to observe the way those specific biomarkers develop through time and if it results in significant differences in the current biomarkers curves.

Regarding the outcomes of GRACE ($\gamma$ and $t+\gamma$ features), we are satisfied with their degree of separability between converters and non-converters. Indeed, we observe that, with the GRACE models, we are able to attain greater prediction accuracy than when using another feature individually, thus extracting a quite reliable disease index that can, upon further validation, be used in clinical environments to perceive the prognosis of a certain patient.

As future work, one idea, not explored in this thesis, is to extract from the final iteration in the construction of the GRACE curves, not only the $\gamma$ parameter for each patient i, but also the different contributions $\gamma_{ij}$ of each biomarker j. With this information, one should be able to perceive the position of each biomarker in the sequence of events, for instance, by ordering the means of each biomarkers $\gamma_{ij}$. Another very interesting idea is to explore this contributions in the scatterplot of $\gamma_i$ of one biomarker against the $\gamma_j$ of another, to detect possible patterns or separation lines between the different classes. This would provide with a progression surface instead of a progression line, which would be constructed from the premise that each biomarker evolves not only in a different moment in time but also at a different rhythm. This idea is further explained in Appendix E, where a picture of that hypothesized 2D plot of $\gamma$'s is shown in Fig.E.1 (Page E-3).

Regarding the classifier results, learning using the disease stage features yielded robust results that follow two distinct patterns for training and for validation. The training results are overall pessimistic,
suggesting that GRACE+ does not yield considerable improvements in weighted AUC from the values attained with the BL. This may be caused by the fact that the features used to represent the baseline dataset in our approach are already established key-features of AD. This limits the potential of GRACE+, since the baseline is already quite optimized, yielding quite good results. Therefore, as future work, the same analyses performed here should be repeated using a larger set of features, including more specific values, such as outcomes of specific questions in neuropsychological tests or more imaging features from specific brain areas.

In other hand, the sensitivity and specificity are quite altered in training when using GRACE+, indicating that with GRACE+ non-converters are being more accurately classified as non-converters while converters are being more often misclassified. For medical purposes, this is not ideal, since the costs of incorrectly classifying a converter surpass the costs of incorrectly classifying a non-converters, but we observe that the increase in specificity often compensates the decrease in sensitivity.

The validation results have demonstrated to be quite positive, since we observe that GRACE+ classifiers, in general, yield results that are closer to the results from training. This means GRACE+ performs an overall much better generalization, which indicates that incorporating features of disease stage mitigates overfitting to the training dataset. The effects of preprocessing in the validation were also analysed and we see that, in general, not only the results with GRACE+ are significantly better than BL, but that those results are often less variant, which suggests that, with GRACE+, the classifiers are less dependent on the employed preprocessing techniques and also less dependent on the dataset.

In summary, the results indicate that for a general data mining approach targeting the detection of future conversion to AD, with a general preprocessing routine and dataset, the incorporation of GRACE+ features is likely to yield improved results. However, future validation is required using independent datasets.
Bibliography


Related Work: Complementary Information
Table A.1: Number of subjects for each diagnosis category on the training dataset used in each of the presented studies: Donohue et al.[5], Young et al.[69], Huang et al.[71], Ewers et al.[82], Zhang et al.[83], Cui et al.[84], Cheng et al.[85], Ye et al.[86] and Barnes et al.[87]. NS=No Subjects; MCIc= MCI converters; MCInc = MCI non-converters.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Amyloid</td>
<td>ApoE+</td>
<td>All</td>
<td>Amyloid</td>
<td>ApoE+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>579</td>
<td>570</td>
<td>285</td>
<td>189</td>
<td>139</td>
<td>347</td>
<td>182</td>
<td>88</td>
<td>350</td>
</tr>
<tr>
<td>NL</td>
<td>100</td>
<td>92</td>
<td>92</td>
<td>34</td>
<td>22</td>
<td>83</td>
<td>101</td>
<td>111</td>
<td>51</td>
</tr>
<tr>
<td>MCI</td>
<td>EMCI: 137</td>
<td>EMCI: 85</td>
<td></td>
<td>MCInc: 72</td>
<td></td>
<td>MCInc: 87</td>
<td>MCInc: 56</td>
<td>MCInc: 177</td>
<td>MCInc: 203</td>
</tr>
<tr>
<td></td>
<td>LMCi: 225</td>
<td>LMCi: 248</td>
<td></td>
<td>LMCi: 58</td>
<td></td>
<td>LMCi: 56</td>
<td>LMCi: 43</td>
<td>LMCi: 142</td>
<td>LMCi: 179</td>
</tr>
<tr>
<td>AD</td>
<td>117</td>
<td>145</td>
<td>64</td>
<td>59</td>
<td>45</td>
<td>81</td>
<td>96</td>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>
Table A.2: Features incorporated in each of the presented studies: Donohue et al.[5], Young et al.[69], Huang et al.[71], Ewers et al.[82], Zhang et al.[83], Cui et al.[84], Cheng et al.[85], Ye et al.[86] and Barnes et al.[87]. 1 includes age, gender, ethnicity, marital status, family history of AD and education; 2 includes history of depression, stroke, hypertension, other cardiovascular diseases, diabetes, respiratory condition, kidney disease, smoking, head injury, thyroid condition; 3 includes low energy, insomnia, abnormal gait, systolic and diastolic blood pressure, pulse, body mass index. For each ROI it was extracted: 4 the cortical thickness, thickness SD, surface area and cortical volume; 5 the cortical thickness, surface area and volume. (T) thickness, (C) used to separate into cohorts, (P) used as the target to predict.
Dataset Description and Statistical Analysis
B.1 Learning Examples - Complementary Information

After computing the learning examples, the resulting dataset (detailed in Table B.1) comprises 1387 instances from 474 distinct ADNI subjects. The number of instances per subject is depicted in Table B.2.

**Table B.1:** Characteristics of the Learning Dataset. For each feature, it is presented the number of available data entries (N), the number of subjects with at least one assessment of that feature (n) and the triplet of 1st, 2nd and 3rd quartiles for each class cohort and the whole combined dataset. It is also presented the t-test p-value, F measure and degrees of freedom, comparing the values from both class cohorts. The majority of features are very separable, with statistically significant differences between converters and non-converters.

<table>
<thead>
<tr>
<th>Feature</th>
<th>N / n</th>
<th>MCInc (n=825 / N = 283)</th>
<th>MCInc (n=562 / N = 222)</th>
<th>Combined (n=1387 / N = 474)</th>
<th>Test - Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMSE</td>
<td>1374/469</td>
<td>25/26/28</td>
<td>27/29/29</td>
<td>26/27/29</td>
<td>P&lt;0.001 343</td>
</tr>
<tr>
<td>CDRSB</td>
<td>1368/469</td>
<td>1.5/2.0/3.0</td>
<td>0.5/1.0/1.5</td>
<td>1.0/1.5/2.5</td>
<td>P&lt;0.001 369</td>
</tr>
<tr>
<td>Hippocampus</td>
<td>1075/438</td>
<td>3.39/3.65/4.25</td>
<td>3.39/4.47/5.01</td>
<td>3.55/4.66/4.6</td>
<td>P&lt;0.001 215</td>
</tr>
<tr>
<td>Entorhinal</td>
<td>1075/438</td>
<td>2578/3004/3535</td>
<td>3276/3738/4216</td>
<td>2762/3324/3920</td>
<td>P&lt;0.001 218</td>
</tr>
<tr>
<td>Whole Brain</td>
<td>1075/438</td>
<td>108/114/119</td>
<td>110/115/120</td>
<td>109/114/120</td>
<td>P&lt;0.02 5.43</td>
</tr>
<tr>
<td>Ventrices</td>
<td>1075/438</td>
<td>19.6/27.2/37.0</td>
<td>15.8/22.1/30.4</td>
<td>18.1/24.7/35.1</td>
<td>P&lt;0.001 36.9</td>
</tr>
<tr>
<td>FAQ</td>
<td>1377/473</td>
<td>2/5/9</td>
<td>0/1/4</td>
<td>1/3/7</td>
<td>P&lt;0.001 331</td>
</tr>
<tr>
<td>ADAS13</td>
<td>1375/473</td>
<td>17.7/21.3/26.0</td>
<td>10.0/13.3/17.7</td>
<td>13.7/18.0/23.3</td>
<td>P&lt;0.001 658</td>
</tr>
<tr>
<td>RAVLT</td>
<td>1381/473</td>
<td>1/2/4</td>
<td>3/4/6</td>
<td>2/3/5</td>
<td>P&lt;0.001 214</td>
</tr>
<tr>
<td>FDG</td>
<td>711/329</td>
<td>5.40/5.79/6.14</td>
<td>6.02/6.34/6.73</td>
<td>5.60/6.06/6.43</td>
<td>P&lt;0.001 202</td>
</tr>
<tr>
<td>TAU</td>
<td>90/46</td>
<td>71/99/135</td>
<td>48/55/81</td>
<td>60/91/130</td>
<td>P&lt;0.001 17.7</td>
</tr>
<tr>
<td>ABETA</td>
<td>90/46</td>
<td>127/140/152</td>
<td>144/159/232</td>
<td>131/144/163</td>
<td>P&lt;0.001 13.1</td>
</tr>
<tr>
<td>PTAU</td>
<td>90/46</td>
<td>27.5/35.0/44.5</td>
<td>17.0/19.0/24.0</td>
<td>21.0/32.5/42.8</td>
<td>P&lt;0.001 18.6</td>
</tr>
<tr>
<td>AV45</td>
<td>202/196</td>
<td>1.28/1.39/1.53</td>
<td>1.01/1.07/1.28</td>
<td>1.05/1.26/1.46</td>
<td>P&lt;0.001 78.2</td>
</tr>
<tr>
<td>PIB</td>
<td>75/43</td>
<td>1.62/2.11/2.23</td>
<td>1.25/1.42/2.06</td>
<td>1.33/1.72/2.18</td>
<td>P&lt;0.002 10.6</td>
</tr>
<tr>
<td>AGE</td>
<td>1374/474</td>
<td>69.6/74.3/79.3</td>
<td>68.3/74.0/79.3</td>
<td>69.1/74.2/79.3</td>
<td>P=0.446 0.58</td>
</tr>
<tr>
<td>SEX: Female</td>
<td>1374/474</td>
<td>39% (324)</td>
<td>32% (198)</td>
<td>36% (502)</td>
<td>P=0.004 Chi-square 8.36</td>
</tr>
<tr>
<td>Education</td>
<td>1374/474</td>
<td>14/16/18</td>
<td>13/16/18</td>
<td>14/16/18</td>
<td>P=0.039 4.28</td>
</tr>
<tr>
<td>APOE4</td>
<td>1374/474</td>
<td>48% (399)</td>
<td>33% (188)</td>
<td>42% (587)</td>
<td></td>
</tr>
</tbody>
</table>

Table B.2: Distribution of subjects according to the number of instance per subject.
Figure B.1: Smooth densities estimates for the set of numeric continuous features and Histograms of discrete features. The distributions result from a smoothing of the actual data to facilitate the visualization of the differences between the two cohorts. Consequently, considering this smooth estimation, density values have to be analysed with caution (for more detailed information, see Table B.1). For instance, as is the case of PIB which due to high missing value percentage presents a smoothed density distribution slightly distinct from the actual distribution. The means of each cohort are represented by vertical dotted lines.

### B.2 Features’ Informative Value

With the purpose of assessing the worth of the attributes in classification, a brief exploratory pre-analysis was performed with feature selection from one training fold with the inclusion of the estimated $\gamma$, using feature ranking methods (Table B.2) and 4 other common feature selection methods implemented in WEKA, such as correlation based Cfs and wrappers using One Rule algorithm, C-4.5 Decision Trees and Naive Bayes. (Fig. B.2).
Figure B.2: Percentage of times each feature was selected from 4 feature different selection methods. The Model features are consistently picked. Regarding the remaining features, ADAS13 shows the best predictive power.

Table B.3: Top 7 selected features in 10-fold CV Gain Ratio and Info Gain rankings. The 2 parameters introduced as features were ranked consistently the first 2 to be picked.
The image contains a page from a document with a table of contents for a section titled "GRACE Outcomes." The table includes two subsections:

<table>
<thead>
<tr>
<th>B.1 Learning Examples - Complementary Information</th>
<th>B-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.2 Features' Informative Value</td>
<td>B-3</td>
</tr>
</tbody>
</table>
C.1 GRACE Iterations

Here, we present the brief analysis performed to assess the convergence of the residual sum of squares (RSS) in the GRACE algorithm with our dataset. In Fig.C.1, we observe the evolution of RSS through the iterations in the model constructed from the training set (75% of the total dataset). The RSS is reduced in a logarithmic pace and seems to stabilize in the 7th iteration, but is further reduced from the 12th iteration. According to the purposes of our study, we opted to use the tolerance employed in [5] of $10^{-3}$, which is reached in the 9th iteration.

![Graph showing convergence analysis](image)

**Figure C.1:** Convergence analysis of the model constructed from the whole training set. The process is stopped by reaching the tolerance level ($10^{-3}$) in the 9th iteration. RSS=Residual Sum of Squares.

In Fig.C.2 and C.3, we present the curves estimated in the iterations prior to the final iteration, which is shown in Fig.4.1.

In the first iteration, we see that the data is still centered and the estimated curves are approximately horizontal. However, we observe that just one iteration results in a considerable change in the shapes of the curves. The curves from iteration 2 present an overall logistic shape, but as we see in Fig.C.1, this actually results in a higher RSS, i.e. further iterations are necessary to better fit the data. In iteration 5, we already observe similar shapes to the final outcomes.
Figure C.2: GRACE progression curves and shifted dataset in iterations 1 and 2, using the training dataset (75% of the whole dataset).
Figure C.3: GRACE progression curves and shifted dataset in iteration 5, using the training dataset (75% of the whole dataset).

Note that for each iteration, the x-axis \((t+\gamma)\) is expanding, which can be observed in the distributions of \(\gamma\) across the different categories, presented in Fig.C.4, C.5, ??. This distributions go accordingly to the expected, obtaining \(\gamma\)'s that represent a quantifiable value of the disease stage, supported by a correct ordering of the different diagnosis categories. The difference between MCI converters and non-converters is the most pronounced, indicating the good separability of this two categories that is attained using \(\gamma\) (as can be perceived in Fig.4.1.1).

Figure C.4: The \(\gamma\) estimations in iteration 1, using the training dataset (75% of the whole dataset).
C.2 GRACE Curves in 10-fold CV

Here, we present the outcomes of the GRACE models constructed from the 50 folds created (5 seeds times 10 folds). As we can see the shapes are similar to the shape obtained using the whole dataset, which was expected since these curves are being created with 90% of that total dataset. The variance is reduced, and is mainly observed in the highly sparse variables, including Aβ, tau, p-tau and PIB. Variance in FDG and Whole Brain is also considerable, which results from the high between-subject variability that we observe in these features.
Figure C.6: GRACE progression curves, from the several CV folds, including the variance obtained from the different datasets.
Classification: Complementary Results

Contents

C.1 GRACE Iterations ................................................................. C-2
C.2 GRACE Curves in 10-fold CV ............................................. C-5
Here, we present some complementary figures from the classification results. Namely, we present the top classifiers using BL (Fig.D.1) and GRACE+ (Fig.D.2). The parameters that were picked from the grid search are presented in Table D.1.

**Figure D.1:** Top 7 classifiers with highest AUCW from the whole set of classifier/preprocessing combinations using the BL dataset. No statistical difference between the top classifier and the other classifiers in the top 6 is observed, being only detected a significant difference with the 7th top classifier.

**Figure D.2:** Top 11 classifiers with highest AUCW from the whole set of classifier/preprocessing combinations using the GRACE+ dataset. No statistical difference between the top classifier and the other classifiers in the top 10 is observed, being only detected a significant difference with the 11th top classifier.
<table>
<thead>
<tr>
<th>Classifier</th>
<th>DT J48</th>
<th>KNN</th>
<th>Logistic Regression</th>
<th>Naive Bayes</th>
<th>Random Forest</th>
<th>SVM Polynomial</th>
<th>SVM Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preprocessing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset</td>
<td>BL</td>
<td>BL</td>
<td>BL</td>
<td>BL</td>
<td>BL</td>
<td>BL</td>
<td>BL</td>
</tr>
<tr>
<td>Optimized Parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confidence</td>
<td>0.15</td>
<td>0.15</td>
<td>×</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Ridge</td>
<td>-5</td>
<td>-9</td>
<td></td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>MinVal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NumTrees</td>
<td>20</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degree Complexity</td>
<td>0</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>2</td>
<td>-1</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMOTE Percentage</td>
<td></td>
<td></td>
<td></td>
<td>0%</td>
<td>50%</td>
<td>75%</td>
<td></td>
</tr>
</tbody>
</table>

Table D.1: Optimized parameters for each classifier's best performance.
Future Work
When closely analysing the underlying processes in the GRACE algorithm, we perceived the assumptions that were made. One assumption is to compute the subject-specific $\gamma_i$ equally influenced by all the biomarker and subject-specific $\gamma_{ij}$'s. This assumption, however, can present a restriction, due to the aspects of the given dataset, such as the high between-subject variability and high missing value percentage of some features. One possible approach would be to weight each biomarker according to its informative value, inversely proportional to the missing values percentage.

An idea, not pursued in this thesis, is to extract each biomarker-specific contribution for each patient. Accordingly, we would have a form of scale transformation from the values of the features or the percentiles values, to a common time-scale (represent by the $\gamma$'s). After that transformation, feature selection would yield a new perspective on the most informative features and classification could detect some new patterns to distinguish MCIc from MCInc. For instance, we could be able to observe that MCIc and MCInc share the same $\gamma$ values for $A_\beta$, which would be expected since this feature is assumed to be the first to become abnormal, but have highly separable $\gamma$ in features such as MMSE or MRI. This potential conclusions could further validate the order of events that is postulated in [3, 4].

Another idea is to plot two different biomarker-specific $\gamma$ (exemplified in Fig.E.1) and see if there it can be detected some way to accurately separate the classes. This form of kernel transformation has the advantage of being interpretable. Furthermore, some knowledge on the insight estimation of GRACE could be extracted from this plots. For instance, Fig.E.1 suggests that RAVLT has an higher power to separate the classes than FAQ, as we observe classes distributions overlapping significantly in the FAQ axis, but considerably more separable in the RAVLT axis.

In other hand, we could detect a pattern demonstrating a shift from the line $\gamma_{RAVLT}=\gamma_{FAQ}$. These deviations would represent that while one of these features is progressing, the other is temporarily stagnant, which would be a very important information that could result in an all new insight on the way AD biomarkers progress. So far, the progression has been modelled in a biomarker event sequence and with such conclusions the result may create evidence that a biomarker can have multiple positions in this sequence, i.e. we could reach the conclusion that a biomarker deteriorates rapidly in an early stage of the disease and then stagnates, to then deteriorate rapidly again, in a future stage.

Note that these are embrionary ideas and further discussion and conceptualization would have to made to pursue these conclusions. Nonetheless, it is important to notice that the spectrum of potential knowledge that can be extracted from the GRACE algorithm goes further than just the approach proposed in this thesis.
Figure E.1: Scatterplot of the estimated $\gamma_{\text{RAVLT}}$ vs $\gamma_{\text{FAQ}}$. 