UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE FÍSICA

A multimodal approach to distinguish MCI-C from MCI-NC subjects

Rochelle Ann Costa Silva

Mestrado Integrado em Engenharia Biomédica e Biofísica

Perfil em Sinais e Imagens Médicas

Dissertação orientada por:

Orientador Externo: Professora Dra. Margarida Silveira, Instituto de Sistemas e Robótica (ISR), Instiuto Superior Técnico (IST)

Orientador Interno: Professor Dr. Nuno Matela, Instituto de Biofísica e Engenharia Biomédica (IBEB), Faculdade de Ciências da Universidade de Lisboa (FCUL)

2016

' Nobody ever figures out what life is all about, and it doesn't matter. Explore the world. Nearly everything is really interesting if you go into it deeply enough.'

- Richard Feynman

Acknowledgements

First of all, I would like to express my very special thanks to Prof. Dr. Margarida Silveira, from Instituto Superior Técnico (IST), for having accepted to be my supervisor for this thesis in Institute for Systems and Robotics (ISR), in Lisbon. I am very grateful for having had the opportunity and privilege to learn and achieve success in one of the most important experience of my academic life. I truly value all her guidance, support, patience in helping to solve the technical problems that occurred, and the knowledge transmitted throughout the development of the present thesis. I also acknowledge all the corrections and comments on this written work, which were fundamental.

I would also like to express my deep gratitude to Dr. Jonathan Young from Centre for Neuroimaging Sciences, King's College London, for responding very promptly to my e-mails and clearly explaining the key steps to implement the method from his article.

I also thank Prof. Dr. Nuno Matela, for having promptly accepted to be my internal supervisor, for being accessible to help me when needed, and for the comments and suggestions on this written work.

My gratitude also goes to Prof. Dr. Eduardo Ducla Soares for presenting the wonderful world of Biomedical Engineering in my 12th grade and to Dr. Maria João Rosa and Dr. Janaina Mourão Miranda with whom I did an internship in London in my third year of the course. They were instrumental in my choice of specialising in such an interesting area which is Machine Learning.

I appreciate all the support and advice provided by my friends, which helped me overcome stressful moments, and thank them for the great memorable times spent together, during these 5 years of my academic life.

Finally, to my beloved parents I thank their love, trust, patience and motivation. A special thank you to my brother Ryan, who has always been there for me at all times.

Abstract

Alzheimer's Disease (AD) is one of the most common neurodegenerative diseases, affecting 60-80% from all dementia cases. Unfortunately, the cure for AD is still not known and only some treatments can be done in its early stages to slow up the symptoms and cognitive decline, avoiding worst patients' living conditions. As most of the AD diagnoses are late, it increases the difficulty of applying the strategies and treatments available. Therefore, current studies aim at detecting AD at an early stage. For this purpose, they are studying mild cognitive impairment (MCI) subjects, as this is normally the first condition before developing AD. Nonetheless, not all MCI patients convert to AD, some remain stable or even may reverse the cognitive decline. In this sense, being able to distinguish between MCI-converters (MCI-C) and MCI-non converters (MCI-NC) reveals a quite important task.

In order to distinguish between these and other groups of subjects many classifiers can be used. Classifiers are machine learning algorithms which apply artificial intelligence. These are extremely useful to identify patterns in, for example, medical brain images, to find disease related patterns and try to achieve an early and reliable diagnosis. The Support Vector Machine (SVM) is a widely used classifier for AD studies and is very appealing as it deals well with high-dimensional problems, which is present when using neuroimages because of the high number of voxels in each image. Nonetheless, SVM is a non-probabilistic classifier and only provides the class predicted for a given test. In a clinical perspective, it would be advantageous to also have a confidence level about the prediction made, to avoid diagnosis being hampered by overconfidence. Hence, of late the interest in probabilistic classifiers is rising. The Logistic Regression (LR) and the Gaussian Process (GP) are examples of probabilistic classifiers, but few studies used these methods to present results for AD classification, additionally the analysis of the posterior probability given by these classifiers is also still not well explored.

In this context, this thesis proposes the comparison of the performance of probabilistic (LR and GP) and non-probabilistic (SVM) classifiers for AD context with special interest in reaching good results for MCI-C vs MCI-NC. These tests were done using two neuroimaging modalities: the deoxyglucose Positron Emission Tomography (FDG-PET) and structural Magnetic Resonance Imaging (sMRI), in single modal and multimodal approach. A whole-brain approach was chosen, to avoid restringing the model just for certain brain regions. For feature selection methods, the LASSO and group LASSO with L_1/L_2 regularization, for both single and multimodality cases, were used respectively. Four different binary classification tests involving AD, MCI and elderly cognitive normal (CN) subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database,

were performed: AD vs CN, AD vs MCI, CN vs MCI and MCI-C vs MCI-NC with a conversion period of 24 months. The results demonstrated the advantage of using GP and LR as they can achieve state-of-the art classification results and be better than SVM, in most cases, while providing posterior probabilities that will help evaluate how confident the classifier is on its predictions. However, to distinguish MCI-C and MCI-NC, SVM seemed to get better results, with LR being just a little worse than SVM. The posterior probabilities from GP attracted more attention, because they demonstrated higher confidence in results, whereas LR posterior probabilities were mostly near the threshold value, meaning that the class is not chosen with a lot of confidence. Although the multimodal approach did not show always the best results, for the MCI-C vs MCI-NC classification it outperformed the single modality results, independently of the classifier used. Thus, exhibits that is useful to joint information of different modalities to help distinguish between MCI-C and MCI-NC.

Keywords

Alzheimer disease (AD), Mild cognitive impairment (MCI), classification, feature selection, posterior probability

Resumo

A Doença de Alzheimer (AD, do inglês *Alzheimer's Disease*) é uma doença neurodegenerativa com crescente prevalência que afecta pessoas com idade mais avançada, habitualmente superior a 65 anos, e constitui entre 60-80% de todos os casos de demência. Provoca uma progressiva degradação dos neurónios e disfunção das sinapses, que constituem a região de ligação entre neurónios. Acredita-se que estas alterações sejam consequentes da acumulação de placas da proteína beta-amyloide no meio extracelular e de alterações anormais na proteína tau no meio intracelular. Consequentemente, com a progressão da doença, o doente começa a manifestar perda de memória, dificuldade em formular pensamentos e alterações do comportamento, chegando a um estado em que se repercute nas atividades da vida diária. Atualmente, não existe cura para a AD, apenas alguns tratamentos que podem ser feitos para tentar retardar os sintomas e o declínio cognitivo. Estes conseguem ser mais eficazes nas primeiras fases da doença evitando assim piores condições de vida para os doentes. Como geralmente o diagnóstico da AD é tardio, a eficácia dos tratamentos disponíveis torna-se ainda mais limitada. Neste contexto, a doença de Alzheimer é vista como um problema de saúde pública com elevado impacto económico, tendo sido identificada como uma prioridade na investigação atual.

Muitos estudos têm como principal objetivo a deteção precoce da AD, para que os tratamentos possam ser usados com a devida antecedência, sendo mais benéficos para o doente. Neste sentido, existe interesse no estudo do défice cognitivo ligeiro (MCI, do inglês: *Mild Cognitive Impairment*), visto que é considerado como um estado prodrómico da doença de Alzheimer, ou seja, doentes com MCI apresentam sintomas que podem indicar o início de AD antes que os sintomas mais específicos da doença surjam. No entanto, nem todos os casos de MCI desenvolvem AD, alguns permanecem estáveis ou podem reverter o declínio cognitivo. Deste modo, tem especial importância conseguir distinguir sujeitos com MCI que poderão converter (MCI-C), num determinado espaço de tempo, dos que não irão desenvolver a doença, ou seja, os MCI não conversores (MCI-NC). Diversos métodos de aprendizagem automática que aplicam algoritmos de inteligência artificial têm sido utilizados para reconhecer padrões nos dados obtidos através de técnicas ou exames médicos. Pretende-se encontrar padrões nos dados relacionados com a doença e alcançar um diagnóstico precoce confiável, através de classificações com elevada precisão obtidas por estes algoritmos. A combinação dos dados médicos com a inteligência artificial deu origem a uma tecnologia interdisciplinar, a que se dá o nome de diagnóstico auxiliado por computador (CAD, do inglês: *Computer-Aided Diagnosis*). Nos exercícios de CAD, em particular quando se usam técnicas de neuroimagem, para a criação um modelo de classificação são definidas normalmente cinco etapas: o pré-processamento das imagens, a extração de características, a seleção

de características, a classificação e a finalmente avaliação do desempenho do classificador. O pré-processamento pode envolver várias fases, sendo essencialmente usado para eliminar a presença de ruído e heterogeneidades e fazer o alinhamento das imagens. Tanto a extracção como a seleção das características permitem reduzir o problema da elevada dimensionalidade existente nas neuroimagens, que advém do excessivo número de voxels/características presentes em cada imagem.

Os exames médicos disponíveis para facilitar o diagnostico da AD são diversos e incluem exames de neuroimagem, análises laboratoriais, testes genéticos e neurofisiológicos. Neste trabalho, foram usadas duas modalidades de imagem que em estudos anteriores provaram ser vantajosas para o diagnostico da AD: a Tomografia por Emissão de Positrões ^{8}F-Fluorodesoxiglucose (FDG-PET, do inglês: *fluorodeoxyglucose Positron Emission Tomography*) que permite detetar hipometabolismo nas regiões afetadas pela doença, e as imagens estruturais de Ressonância Magnética (sMRI, do inglês *structural Magnetic Resonance Imaging*) que permitem detetar perda de volume do tecido cerebral. Ao juntar a informação destas duas modalidades, é possível fornecer ao classificador diferentes tipos de informação, funcional e estrutural, podendo alcançar previsões mais precisas. Por conseguinte, estas técnicas foram testadas individualmente, mas também numa abordagem multimodal.

Para evitar o elevado número de voxels/características presentes nas imagens, determinados estudos usam apenas certas regiões do cérebro. No entanto, foi preferida a abordagem em que todos os voxels/características do cérebro são usados para não limitar o estudo apenas a determinadas zonas. Para selecionar as regiões mais relevantes de todo o cérebro e diminuir o problema da dimensionalidade foram usados dois métodos de seleção de características: o LASSO, para o caso em que se usou cada modalidade individualmente, e o group LASSO multi-task, no caso multimodal.

O classificador mais utilizado para estudos de AD é a máquina de vetores de suporte (SVM, do inglês: *Support Vector Machine*). Este classificador é apelativo por se adequar a problemas de elevada dimensionalidade e apresentar bons resultados. No entanto, SVM é um classificador não-probabilístico, ou seja, devolve apenas a classe que prevê para um determinado teste e não uma probabilidade associada. Numa perspectiva clínica, seria mais vantajoso ter uma medida de confiança quanto à previsão feita pelo classificador. Recentemente, foram introduzidos dois classificadores que devolvem probabilidades à posteriori: o Processo Gaussiano (GP, do inglês *Gaussian Process*) e a Regressão logística (LR, do inglês *Logistic Regression*). Porém, ainda não foram muito explorados em estudos de AD, especialmente em relação às suas probabilidades à posteriori.

Neste âmbito, com a presente tese testaram-se três classificadores (SVM, GP e LR), numa perspectiva multimodal, que junta dados FDG-PET e sMRI da base de dados

Alzheimer's Disease Neuroimaging Initiative (ADNI), bem como numa abordagem usando as modalidades individualmente. Estes classificadores foram utilizados em quatro testes de classificação diferentes, nomeadamente, para distinguir: AD de sujeitos com idades avançadas e cognição normal (CN); AD de MCI; CN de MCI e com maior interesse os MCI-C de MCI-NC, num período de tempo de conversão 24 meses. A partir dos resultados obtidos foi possível verificar que tanto o GP como o LR apresentaram resultados de classificação melhores que o SVM, para os casos AD vs CN, AD vs MCI e CN vs MCI. No entanto, na classificação verdadeiramente pertinente em termos científicos, ou seja, quando se testou MCI-C vs MCI-NC, o SVM revelou melhores resultados, sendo que o LR não ficou muito abaixo do SVM, já o GP teve uma performance inferior. É importante salientar que o GP apresentou vantagens em relação às probabilidades à posteriori exibidas pelo LR, visto que demonstrou mais confiança nas previsões feitas, enquanto o LR apresentou probabilidades à posteriori mais próximas do limiar entre a escolha de pertencer a uma classe ou outra. Com esta diferença foi possível demostrar a relevância de ter em consideração a análise das probabilidades à posteriori, em vez de se limitar à analise da precisão do classificador. Em relação ao número de características usadas, o LR necessitou um maior número em comparação ao GP ou SVM, apesar disso, não revelou ter um custo computacional superior aos outros dois classificadores. Quanto aos métodos de seleção de características, LASSO e group LASSO multi-task, destaca-se que ambos foram eficientes em diminuir o número de características e selecionaram regiões pertinentes, como o hipocampo, amigdala, tálamo, putamen e ventrículo lateral, que estão de acordo com as regiões detectadas em estudos anteriores. Em alguns casos, a abordagem multimodal não revelou ser superior aos resultados obtidos usando as modalidades individualmente. Não obstante, para a distinção entre MCI-C vs MCI-NC, independentemente do classificador usado, os resultados foram melhores aos obtidos quando se usou as modalidades individualmente. Assim demonstra-se que uma abordagem multimodal apresenta vantagens para diferenciar estes dois grupos de sujeitos.

Palavras Chave

Doença de Alzheimer, Défice cognitivo ligeiro, classificadores, selecção de características, probabilidade à poteriori

Contents

List of Figures

List of Tables

Acronyms

AD	**A**lzheimer's **D**isease
ADNI	**A**lzheimer's **D**isease **N**euroimaging **I**nitiative
a**MCI**	amnestic **M**ild **C**ognitive **I**mpairment
APP	**A**myloid **P**recursor **P**rotein
APOE	**A**polipoprotein E
Aβ	Protein beta-**A**myloid
CAD	**C**omputer **A**ided **D**iagnosis
CAP	**C**omposite **A**bsolute **P**enalties
CN	**C**ognitive **N**ormal
CSF	**C**erebro**S**pinal **F**luid
DPC	**D**ata sharing and **P**ublications **C**ommitte
FAQ	**F**unctional **A**ctivities **Q**uestionnaire
FDG-PET	^{18}F-**F**luoro-**D**eoxy-**G**lucose **P**ositron **E**mission **T**omography
FN	**F**alse **N**egatives
FP	**F**alse **P**ositives
GM	**G**ray **M**atter
GP	**G**aussian **P**rocess
GPML	**G**aussian **P**rocesses for **M**achine **L**earning
ICP	**I**shihara **C**olor **P**late
IDA	**I**mage **D**ata **A**rchive
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
LIBSVM	**L**ibrary for **S**upport **V**ector **M**achines
LONI	**L**aboratory of **N**euro **I**maging
LR	**L**ogistic **R**egression
MCI	**M**ild **C**ognitive **I**mpairment

MMSE	**M**ini-**M**ental **S**tatus **E**xam
MNI	**M**ontreal **N**eurological **I**nstitute
MRI	**M**agnetic **R**esonance **I**maging
NIA	**N**ational **I**nstitute on **A**ging
NINCDS-ADRDA	**N**ational **I**nstitute of **N**eurological and **C**ommunicative **D**isorders and **S**troke-**A**lzheimer's **D**isease and **R**elated **D**isorders **A**ssociation
PAD	**P**re-dementia **A**lzheimer **D**isease
PIB	**P**ittsburgh **C**ompound **B**
ROI	**R**egion **O**f **I**nterest
SLEP	**S**parse **L**earning with **E**fficient **P**rojections
SPM	**S**tatistical **P**arametric **M**apping
SVM	**S**upport **V**ector **M**achine
TN	**T**rue **N**egatives
TP	**T**rue **P**ositives
WM	**W**hite **M**atter
WHO	**W**orld **H**ealth **O**rganization

Symbols

Greek Symbols

α	Lagrange multiplier
λ	Regularization parameter
$\lambda_{\max}$	Maximum regularization parameter defined by SLEP toolbox
θ	Hyperparameter
θ_{MRI}	Hyperparameter which defines the weight for MRI data
θ_{PET}	Hyperparameter which defines the weight for PET data
ξ_n	Slack variable

Roman Symbols

$\mathbf{K}$	Kernel matrix
$\mathbf{W}$	Feature weight matrix
$\mathbf{w}$	Feature weight vector
$\mathbf{X}$	Feature data matrix
$\mathbf{x}$	Feature data vector
$\mathbf{X}_{\text{MRI}}$	Feature matrix for MRI data
$\mathbf{X}_{\text{PET}}$	Feature matrix for PET data
b	Bias term
C	SVM parameter

D	Data set
d	Number of features
L	Lagrangian
n	Number of samples/subjects
s	Weight for a given sample
t	Number of tasks/modalities
y	Sample label, (positive +1, or negative -1)

Chapter 1

Introduction

Alzheimer's Disease (AD) is one of the most common neurodegenerative disorders in older people, accounting for 60-80% of age-related dementia cases (Ye et al., 2011). The disease causes neurons progressive damage or destruction and loss of their connections in the brain, consequently the patient begins losing memory, thinking and behavior abilities and reaches a state that they are entirely unable to take care of themselves, having difficulties controlling even the most basic necessities and consequently, require around-the-clock care. Most often, AD is diagnosed in people over 65 years of age (late-onset), however some individuals younger than age 65 (early-onset) can also develop the disease, but the risk of getting this disease is much higher as people get older. Unfortunately, as neurons normally are not able to regenerate and do not undergo cell division, all the caused damage in the brain cannot be recovered. Therefore, till date AD is considered an irreversible brain disease which leads ultimately to death, because there is currently no known cure and present treatments cannot stop AD from progressing, they only can slow down the worsening of symptoms. The speed of progression can vary, but an average point for survival time ranges from 3.3 to 11.7 years, with most cases in the 7 to 10-year period (Todd et al., 2013).

Brookmeyer in 2007 reported that there were 26.6 million cases of AD in the world in 2006 and in 2050 one person in 85 will suffer from AD (1.2% of total population) or 106.8 million (Cornutiu, 2015). In Portugal the numbers from 2012 indicate that more than 182 000 people suffer with dementia (this represents 1.71% of the population, a little higher than the European mean which is 1.55%) (Alzheimer Europe, 2013). A

report estimates from (World Health Organization (WHO) and Alzheimer's Disease International, 2012) point that the numbers from 2012 will double until 2030, and more than triple by 2050. As most of the AD diagnoses are late, it increases the difficulty of applying some strategies which are used actually to try reducing the progression of the disease. For this reason and due to its big emotional and financial impact on society, AD is a quite concerning public health issue and has been identified as a research priority (Ballard et al., 2011).

Patients suffering from AD at a prodromal stage, i.e. when early symptoms appear and might indicate the start of the disease before the characteristic symptoms occur, are, mostly, clinically classified as amnestic mild cognitive impairment (aMCI). When referring to a patient with MCI it means that the patient has an early loss of brain function before meeting criteria for the diagnosis of dementia. In most cases, the function lost is memory, thus, commonly it can be named as aMCI. These patients show cognitive decline greater than expected for their age and education level, however these alterations are not severe enough to interfere with everyday activities (Alzheimer's Association, 2016). According to (Petersen et al., 2010) study, they suggested that about 16% of elderly people with no dementia are affected by MCI and that approximately two-thirds of those with MCI have aMCI. Studies have also compared the rate of conversion of MCI, they have shown that MCI patients convert to AD at an annual rate of 10-15% per year compared with healthy controls who develop dementia at a rate of 1–2% per year (Bischkopf et al., 2002). So, older MCI patients are at a greater risk of developing AD. The patients that do indeed convert to AD are named as MCI-converters (MCI-C). However, not all MCI patients will develop AD, some either develop other forms of dementia (Vascular dementia; Dementia with Lewy bodies; Parkinson's disease; Huntington's disease), remain stable, or in a small minority, revert the process and go back to normal cognition, so these are seen as MCI-non converters (MCI-NC), figure 1.1 describes this division. It is unclear why some MCI patients develop AD or other dementia and others do not (Alzheimer's Association, 2016).

1.1 Alzheimer's Disease

AD is named after the German physician Dr. Alois Alzheimer, who first described this disease in 1906 (Hippius and Neundörfer, 2003). He detected a dramatic shrinkage of the

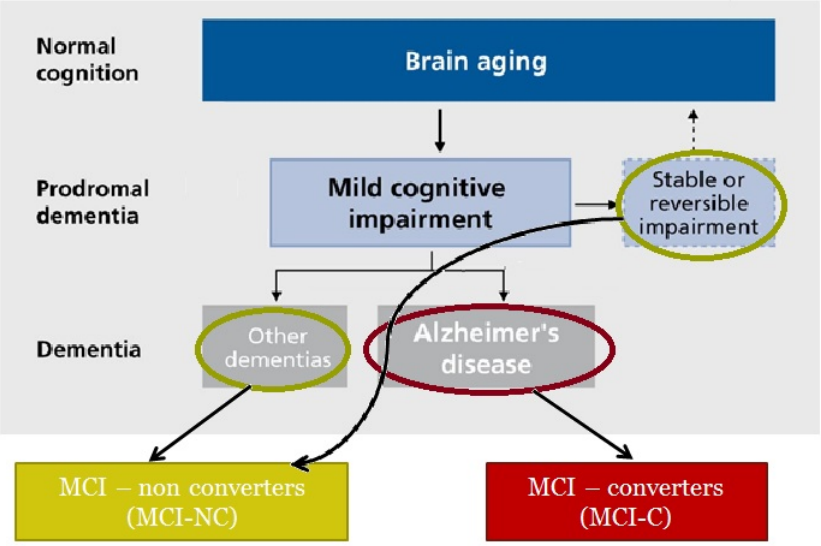


Figure 1.1: A conceptual model of possible developments after reaching MCI state. They can convert to AD (MCI-C) or not (MCI-NC). Adapted from (Golomb et al., 2004)

brain and abnormal deposits in and around nerve cells when analysing the autopsy of a patient who had profound memory loss and many psychological changes. Since then, scientist have been investigating how AD affects the brain, trying to understand its real cause and also several efforts to know how it can be treated are being made, but still with little or no success.

In a healthy adult brain there are around 100 billion nerve cells (neurons), which are the core structural and functional components of the brain and the nervous system. Typically the structure of a neuron consists of dendrites which receive the neural signal, a cell body that will process the signal and an axon which will pass the signal electrically through the neuron and when the electrical signal reaches the end of the axon this causes the terminal branches to release chemical messengers called neurotransmitters. Therefore, the neural communication actually involves an electrochemical communication. An example of a neuron is presented in figure 1.2. In turn, these cells can connect to each other by spaces called synapses, which count for approximately 100 trillion (Alzheimer's Association, 2016). The neurotransmitters travel across these synaptic clefts and bind to the receptors present in the dendrites from neighbour neuron's. This transmission will cause the other neuron to become electrically active and the same process continues and passes through other neurons.

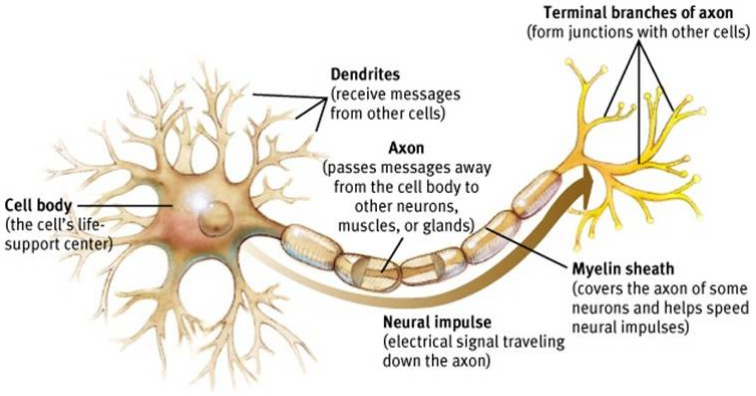


Figure 1.2: Structure of a neuron and how the nerve impulse travels. (http://www.appsychology.com/Book/Biological/neuroscience.htm)

The exact cause of AD is still to be fully understood. However, based on several research done for AD along these years, two pathological hallmarks are known: the accumulation of plaques of the protein beta-amyloid (Aβ) outside neurons and the formation of an abnormal form of the protein tau (neurofibrillary tangles) inside the neurons (Ballard et al., 2011). See figure 1.3 for illustration.

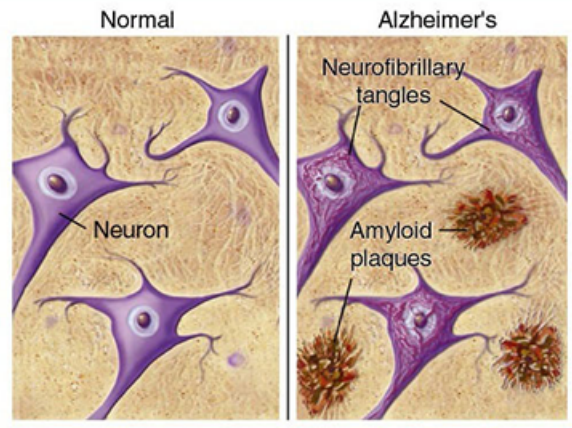


Figure 1.3: Illustration of the two pathological hallmarks in AD: formation of amyloid plaques betwen neurons and neurofibrillary tangles inside the neurons. (http://www.brightfocus.org/alzheimers/)

The first hypothesis about the amyloid plaques suggested that the total amyloid load had a toxic effect on neurons and consequently lead to neurons failure. With more studies in this area, the pathological changes were more deeply investigated, more precisely the

(Aβ) processing, and a more detailed hypothesis was formulated: the amyloid cascade hypothesis. According to this hypothesis, the protein (Aβ) results from the cleavage of the amyloid precursor protein (APP) and accumulates inside neuronal cells but also extracellulaly where it aggregates into plaques and is believed to interfere with the neurons communication by causing synaptic dysfunction and neuronal death (figure 1.4).

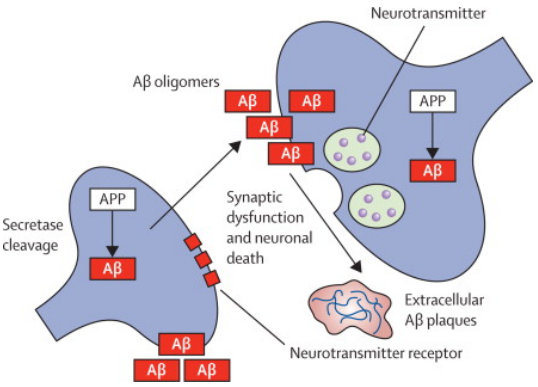


Figure 1.4: Illustration of the amyloid cascade hypothesis. Illustration from (Ballard et al., 2011).

The tau tangles are believed to unable the transport of nutrients and other essential molecules inside neurons contributing therefore for their death.

These processes and changes in the brain are progressive. The first changes can occur without the patient feeling it (clinically silent), as the brain tries to compensate the caused damages (Alzheimer's Association, 2016). However, as the process starts evolving, the brain can no longer compensate the damages and the first symptoms start appearing in accordance to the brain regions affected. A scheme of the affected areas along AD progress and the symptoms are presented in figure 1.5.

The first areas affected are normally in the medial temporal lobe from the brain cortex, which suffers a shrinkage. More specifically the hippocampus is quite affected. As this part of the brain plays a key role in formation of new memories, patients start having difficulties in storing short-term memories. As years pass more neurons are affected from other brain areas like lateral temporal and parietal temporal lobes; consequently the patient begins suffering some difficulties in other activities like reading and recognising objects. The disease spreads also to the frontal lobe. The frontal lobe is responsible for executive functions such as planning, judgment, decision-making skills, and attention.

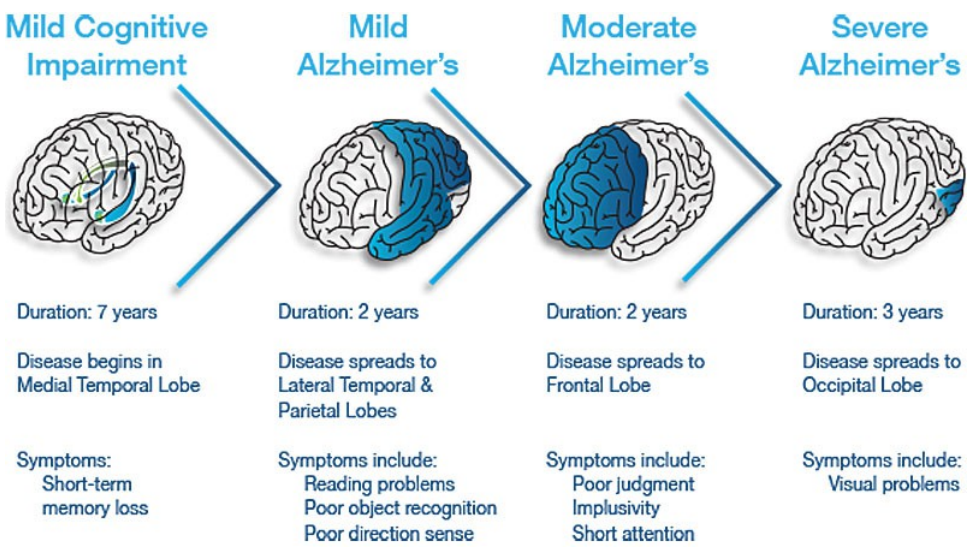


Figure 1.5: Changes in the brain along AD progression and the respective loss of patient capabilities. (http://my-dementia.co.uk/Stages%20and%20Cases.html)

Consequently, in this phase, all these functions can decrease drastically. In a more severe stage the disease reaches the occipital lobe and difficulties in seeing clearly rise.

The diagnosis of AD, at present, can only be done with certainty in autopsy by performing a histopathological confirmation, which involves a microscopic examination of brain tissue. Thus, clinically, only probable diagnosis is possible. In addition, as this disease is quite complex and not fully understood, a single medical test will not be sufficient. The diagnosis has to be carefully evaluated by a physician, normally along with a neurologist help, by following some established guidelines. In this context, the physician can require many different tests and patients' family help, in particular, as explained in (Alzheimer's Association, 2016), these include: 1- Obtaining medical background (including psychiatric and cognitive history) and family history from the patient; 2- Requesting a family member or an person close to the patient to describe the changes in thinking skills and behavior; 3- Executing cognitive tests and physical and neurologic examinations; and 4- Acquiring patients' blood tests and brain images.

1.2 Motivation and problem identification

There is a strong belief that pathological manifestations of AD may appear around 20 or more years before subjects become symptomatic (Alzheimer's Association, 2016). Therefore, it is important to find a way to diagnose even before the classical symptoms

appear. An early and accurate diagnosis will allow patients to benefit from new treatments or strategies that may delay the progress of the disease. In this sense, the aim of today's investigations in this area, is mainly to find the best possible methods which will distinguish between MCI-C and MCI-NC, in order to know which patients will develop AD and need treatment in a near future (i.e. within a few years), and target the disease before irreversible damage or mental decline has occurred.

However, unfortunately the task of predicting conversion from MCI to AD is still known to be difficult and presents challenges beyond that of classifying AD and cognitive normal (CN) subjects or even that of classifying AD/CN vs MCI subjects. For that reason many studies have achieved good results distinguishing AD from CN or CN from MCI, but studies which analyzed MCI-C vs MCI-NC still have low classification performances. This difficulty may be due to the "lag" between brain atrophy and cognitive decline (Hinrichs et al., 2011).

From a public health perspective, treatments as well as clinical trials of therapeutics are classified in terms of primary, secondary, and tertiary prevention interventions (Cavedo et al., 2014). Primary prevention aims at reducing the incidence of illness across the broad population by treating the subjects before disease appears, in other words, it tries to eliminate the potential causes of the disease. Secondary prevention aims at preventing disease at preclinical phases of illness. While tertiary prevention is focused on treating the disease when it has been clinically diagnosed. In AD context, is seems obvious that the primary and the secondary preventions are the ones which concern the population because there is still no known cure for AD, and so tertiary prevention interventions are still not very useful. In case of AD, when referring to a primary prevention it means distinguishing between healthy and MCI patients, when referring to a secondary prevention would be referring to detecting MCI subjects which will convert to AD.

1.3 Use of machine learning for early diagnosis

Machine learning is a subfield of computer science which uses algorithms of artificial intelligence to perform pattern recognition, i.e., to identify patterns and regularities in the data in order to build a model that will make accurate predictions on new data.

Recent advances in neuroimaging techniques and image analysis have significantly contributed to better understand the factors which change the brain and are associated with Alzheimer's disease. Combining them with machine learning algorithms will bring enormous help in finding the best process to reach an early diagnosis of this disease. This combination of elements of artificial intelligence and digital image processing gave rise to a relatively young interdisciplinary technology called computer-aided diagnosis (CAD). For this reason, CAD has gained increasing attention in the medical field in order to simplify the task of interpreting test results by constructing a set of algorithms and computational techniques which use pattern recognition to make future predictions and correctly classify a certain patient.

To define a good classification model when using imaging data, five key steps need to be followed: pre-processing of the data, feature extraction, feature selection, classification and finally the evaluation of the performance of the classification results.

The first steps: pre-processing and feature extraction and feature selection are crucial to perform when using neuroimaging data because without these most probably good results for classification would be difficult to get. This is because medical images can have noise and intensity-inhomogeneity, and in addition, when comparing different scans they might not all be aligned so pre-processing will overcome these issues. The pre-processing can include: motion correction with realignment, spatial normalization and spatial smoothing. Furthermore, when dealing with CAD, there exists the high dimensional problem of neuroimaging data, as neuroimages are characterized by having high dimensionality, i.e. having a very large number of voxels in each image. Thus, when analyzing pattern recognition for neuroimaging studies, the number of voxels is much higher than the number of scans/subjects available. This leads to two big problems: it will require a large amount of memory and computation time and can lead to overfitting, which means that it gives rise to a model that overfits the training sample and generalizes poorly to new samples. One of the first steps to overcome this issue is by performing feature extraction. This can be done by extracting, for example, some predefined regions of interest (ROI) which can be meaningful for the study. A feature selection on the extracted features can then be performed. Feature selection is the process of selecting a subset of features that can be meaningful and relevant for the classification procedure. The feature selection is quite crucial for various reasons: it simplifies the model facilitating interpretation; it makes it computationally more effective as it requires

shorter training times and it enhances generalization by reducing overfitting. Hence, an effective feature selection could not only speed up computation, but also improve the classification performance (Liu et al., 2014).

It is also very common to use kernel methods to solve the high dimensionality of image data. Kernel methods consist of a collection of algorithms based on pair-wise similarity measures between all examples (feature vectors), summarized in a kernel matrix that will have $n \times n$ dimensions instead of data matrix dimensions $n \times d$ (n-number of subjects/scans; d- number of voxels). Given two feature vectors, a kernel function returns a real number characterizing their similarity. The simplest operation one can perform to measure the similarity between two vectors is a dot product (linear kernel). So, in pattern recognition the kernel matrix is many times used instead of the data matrix to simplify the calculations when using neuroimages data, because it makes the model computationally more efficient.

There are many types of machine learning algorithms like supervised learning, unsupervised learning and semi-supervised. Unlike in unsupervised learning, in supervised learning all the training data (i.e. the examples provided to the classifier) are properly labeled by hand (i.e. all examples have their class). It consists in two important steps: training and testing. During the training phase, the algorithm learns some mapping between patterns and the labels and then creates a function that can accurately predict the labels for unseen new patterns. For AD classification problems, the method mostly used is the supervised learning method. Nonetheless, semi-supervised learning algorithms are also recently being tested. Furthermore, in supervised learning we can have two types of pattern recognition: classification and regression. If one wants to distinguish classes or subjects (e.g. distinguish MCI from AD) the learnt function is a classifier model and the labels are discrete values, for example -1, 1 for negative class and positive class, respectively. If instead, one wants to predict a specific value (e.g. values of cognitive test scores), then the learnt function will be a regression model and the labels in this case are continuous values.

The final step, after using the classification algorithm, is then to validate it by evaluating its performance. This evaluation is done by performing a cross-validation and looking into the statistics of the model. A good classification model, will present high accuracy (test's ability to correctly detect or exclude a condition correctly), high sensitivity (test's

ability to identify a condition correctly) and high specificity (test's ability to exclude a condition correctly) values. These statistical measurements are calculated by the formulas presented in figure 1.6, where TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the false negatives. This figure 1.6 presents the machine learning framework usually followed for classification problems using neuroimaging data.

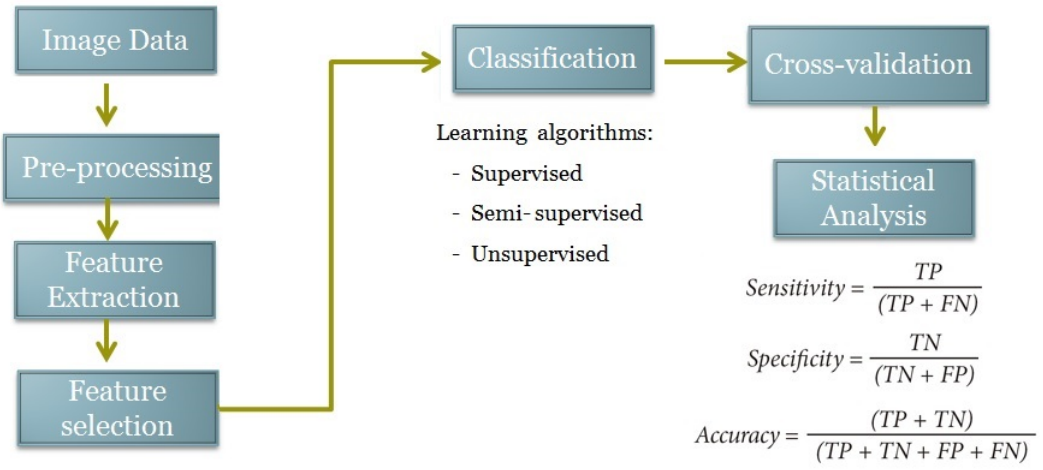


Figure 1.6: Representation of the framework for neuroimaging classification problems.

For AD classification studies, a wide range of different medical information could be used as data to distinguish between patients with AD from those who do not have the disease, or to determinate which are more likely to develop AD. These data, in a machine learning context, are normally called modalities, and include: neuroimages, neuropsychological tests, genetic tests, and cerebrospinal fluid (CSF) tests. The two most widely used neuroimage modalities are: ^{18}F deoxyglucose Positron Emission Tomography (FDG-PET) that measures the cerebral metabolic rate for glucose, and structural Magnetic Resonance Imaging (sMRI) which provides information about brain morphometry and therefore can capture brain tissue atrophy related to the loss of neurons (Vemuri and Jack, 2010). Furthermore, these could be used separately or together giving rise to a multimodal classification approach. Existing studies have indicated that different modalities can provide essential complementary information and therefore improve accuracy in disease diagnosis, some of these studies will be presented in chapter 2. An illustration of these two image modalities is presented in figures 1.7 and 1.8.

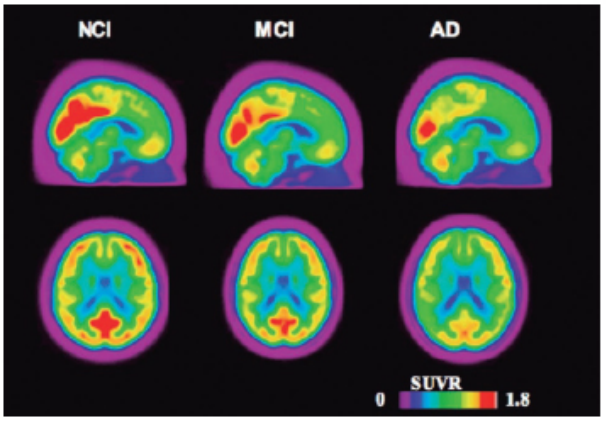


Figure 1.7: Example of FGD-PET images from cognitive normal (left), MCI (middle) and AD (right) subjects. The color scale represents the magnitude of ^{18}F-FDG standardized uptake value ratio (SUVR) which is proportional to glucose uptake. By these images one can identify lower SUVRs in MCI and AD when compared to cognitive normal. Illustration from (Schilling et al., 2016).

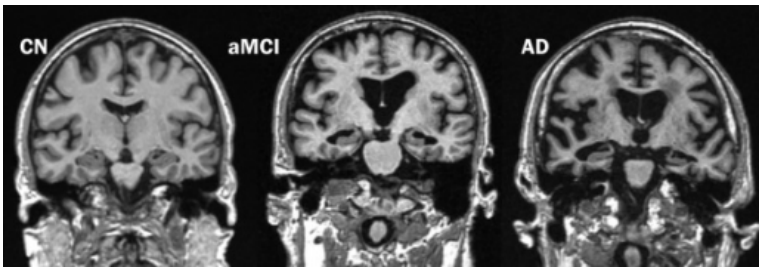


Figure 1.8: Structural MRI images from an older cognitively normal (left), an amnestic mild cognitive impairment (middle) and an Alzheimer's disease (right) subjects demonstrating progressive brain tissue atrophy. Illustration from (Vemuri and Jack, 2010).

The feature selection for the neuroimage multimodal case can be performed independently in each modality. However, this may overloook the complementary information conveyed in different modalities. More recently, studies have used multi-task learning to perform the feature selection for a multimodal approach, where each modality is seen as a task. Multi-task learning is based on a procedure which takes a number of tasks simultaneously and exploits the commonalities between them. So the objective is to defect the intrinsic relationship among different tasks, which can lead to a better results than when learning the tasks independently (Liu et al., 2014).

In terms of classifiers, these can also be distinguished by their property of providing a probability. Most of the classification done in AD studies are based in non-probabilistic classifiers (Zhang et al., 2011; Hinrichs et al., 2011; Liu et al., 2014; Jie et al., 2015), which only provide the class that a sample should belong to. Nonetheless, interest in

probabilistic classifiers has been recently presented (Young et al., 2013; Challis et al., 2015) as these may be advantageous in terms of clinical use because they provide additional information, more precisely, a measure of confidence of the prediction made.

1.4 Contribution of Thesis and Thesis Outline

Considering that classification methodologies to distinguish MCI-C from MCI-NC are still in progress (Golomb et al., 2004; Hinrichs et al., 2011; Davatzikos et al., 2011; Young et al., 2013; Zhang et al., 2014; Cheng et al., 2015; Jie et al., 2015), and taking into account the recent interest in probabilistic classifiers (Young et al., 2013; Challis et al., 2015), this thesis proposes the analysis of a multimodal approach to distinguish these two group subjects using non-probabilistic and probabilistic classifiers. More precisely, in this multimodal procedure two modalities will be used as data: MRI and FDG-PET images. For feature selection the group LASSO multi-task feature selection method provided by (Liu et al., 2009.) software will be used to jointly select features between the two modalities from a whole-brain problem. This whole-brain approach is chosen in order to give the possibility of finding new regions of the brain which can be seen as relevant features, instead of using just predefined regions. For the classification step, three classifiers will be studied: the Support Vector Machine (SVM) which is the most widely used classifier for neuroimaging studies and represents a non-probabilistic classifier; and two probabilistic classifiers, which are not so widely used as the first one, the logistic regression (LR) and the Gaussian Process (GP). In this thesis essentially four analyses will discussed: the analysis of the feature selection step to achieve the best performance for each classifier, the comparison of the performance of these classifiers, the interpretation of the posterior probabilities given by LR and GP, and the investigation of the selected brain regions.

In chapter 2 the State-of-the-art will be presented. The theory of the methods used and their respective toolbox is explained in chapter 3. In chapter 4 all the performed work in this thesis is depicted. This chapter has three main sections: section 4.1 which starts by describing the data used in this work, section 4.2 which shows the experimental design followed to implement the methods proposed and section 4.3 that presents the results and discussion. All the results in this last section are for 4 groups: AD vs CN, AD vs MCI, CN vs MCI and MCI-C vs MCI-NC. Each of these for single modalities (PET

and MRI) and for multimodality (PET + MRI). Finally, the conclusion and future work suggestions are presented in chapter 5.

The innovation in this thesis is the comparison of the performance of non-probabilistic and probabilistic classifiers in a whole-brain approach using LASSO with L_1 regularization and group LASSO with L_1/L_2 regularization as the feature selection methods, for both single and multimodality respectively, and the analyses of the posterior probabilities provided by LR and GP.

Chapter 2

State-of-the-art

The publication of the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria in 1984 represented a breakthrough in the diagnosis of AD (McKhann et al., 1984). These criteria established that the clinical diagnosis should be based on certain characteristics: medical history, clinical examination, neuropsychological testing, and laboratory assessments. However, most demented patients were seen by community physicians who often did not detect dementia or misdiagnosed it, and for pre-dementia AD (PAD) detection this criteria did not show enough sensitivity (Alom et al., 2012). In this context, researchers began developing studies which test machine learning methods in order to understand if new guidelines could be suggested to facilitate physicians in the diagnosis process. Presently, with the research done since then, these guidelines from the 1984 criteria were updated with new guidelines established by the National Institute on Aging (NIA) and the Alzheimer Association, in 2011 (Sperling et al., 2011; Albert et al., 2011; McKhann et al., 2011; Jack et al., 2011). According to this criteria, the brain changes due to Alzheimer's begin several years (20 or more) before symptoms, and suggest that, in some cases, MCI is an early stage of Alzheimer's, whereas the earlier criteria from 1984 would require the appearance of memory loss and cognitive decline in order to make a diagnosis of AD. Although this new criteria for a preclinical phase of AD is still not used by doctors for clinical diagnosis, it will help a lot for research purposes. These new guidelines also added other tests that could facilitate an early diagnosis, like for example the use of neuroimages.

One of the first studies using machine learning for AD diagnosis was (Shankle et al., 1997). They used machine learning algorithms combined with clinical data such as subjects demographic data (age, gender, education) and neuropathological tests, like Functional Activities Questionnaire (FAQ), the Mini-Mental Status Exam (MMSE), and the Ishihara Color Plate (ICP) tasks, to learn rule sets that would help detect very early stages of dementia from normal aging and the results were as good as or better than any rules derived from knowledge provided by expert clinicians. This demonstrated the importance of exploiting machine learning techniques for AD early diagnosis.

Most machine learning studies in this area are making efforts to understand which biomarkers are able to indicate early stages of AD. A biomarker is a biological factor that can be measured to accurately and reliably indicate the presence or absence of disease, or risk of developing a disease (Alzheimer's Association, 2016). Examples being studied for AD include beta-amyloid and tau levels in CSF, genetic risk factors and brain changes detectable by brain imaging techniques. However, the difficulty rises as these indicators may change at different stages of the disease process. In addition, for a biomarker to be validated in order to be used as medical test for clinical diagnosis, multiple studies in large groups of people have to be made and proven that it accurately and reliably indicates the presence or absence of AD. These factors make it difficult to validate biomarkers for AD and researchers are still investigating these promising candidates. Although these candidates are not still seen as entirely validated biomarkers for clinical use, they can provide comprehensive information about the disease being helpful to investigators in machine learning studies and also to physicians and neurologists, specially the neuroimaging techniques, as they also allow the detection of brain changes associated with AD. These techniques will be presented in the following paragraphs.

Structural Magnetic Resonance Imaging (sMRI) is an extremely important image modality for AD studies since it helps detecting brain atrophy as it was presented in previous chapter in figure 1.8. In AD, brain atrophy occurs in a characteristic topographic distribution, it begins in the medial temporal lobe and spreads to the lateral temporal areas, and medial and lateral parietal areas (Cavedo et al., 2014). The most common sMRI measure employed in AD is the atrophy of the hippocampus, recently recommended by the revised criteria for AD as one of the core biomarkers (Albert et al., 2011; McKhann et al., 2011; Jack et al., 2011). In this sense, several studies proposed classification methods to discriminate between patients with AD or MCI and CN based on sMRI like

for example (Klöppel et al., 2008; Davatzikos et al., 2008). Comparing performance results of these studies an other sMRI studies may not be fully correct because they were assessed on different populations. Many factors as: degree of impairment, age, gender, genotype and level of education could be affecting the evaluation of the prediction accuracy. Therefore, in order to have the possibility to compare different performance results (Cuingnet et al., 2011) studied 10 methods of classification of AD using sMRI with the same population. Most of the methods showed high accuracy in classifying AD and CN, nevertheless at the prodromal stage, i.e. for MCI, their sensitivity was very low. This suggests the need of combination with other modalities, to overcome this difficulty.

Functional Magnetic Resonance Imaging (fMRI) has also been used for AD detection. This modality tracks changes associated with blood flow, more precisely, it measures the blood-oxygen-level dependent (BOLD) signal, reflecting the regional neuronal activation and intracortical processing. As a primary prevention biomarker, it still needs considerable research and development work, because it has the issue of possible confound between normal aging and development of AD-related pathology. Normal aging alters potential fMRI biomarker and alterations that are seen in MCI group are similar to middle aged healthy controls (Cavedo et al., 2014). Therefore, at the moment, the main use of fMRI would be in secondary prevention trials, which is actually the main concern in this area. However, it should be noticed that pathologies other than AD, such as Major Depression, can mimic the symptoms experienced by MCI patients, and have also been shown to induce changes in functional connectivity (Challis et al., 2015). So, further work needs to be done to better understand the relationship between BOLD signal and clinical changes in dementia and non dementia cases.

As already referred in chapter 1 and presented in figure 1.7, another imaging modality widely used for AD detection and study of its progression is ^{18}F-fluorodeoxyglucose (FDG) positron emission tomography (PET). FDG-PET is a functional marker that measures tissue uptake of glucose and can therefore be linked to the detection of cortical synaptic dysfunction. Normally it can reveal hypometabolism in the temporoparietal regions, posterior cingulate cortex, and frontal lobe even prior to atrophy (Schilling et al., 2016). This image modality has been approved in the USA for diagnostic purposes and is sensitive and specific for AD detection in its early stages (Ballard et al., 2011) and revealed to be a good predictor of MCI progression within the next 2 years combining clinical covariates (Shaffer et al., 2013).

In recent years, researches have also studied a new modality called carbon 11-labeled Pittsburgh Compound B (^{11}C-PIB) PET. This modality provides both perfusion and amyloid deposition information. Combining this modality with structural MRI good results were achieved (Liu et al., 2015). However, this study was basically to distinguish between AD vs CN (accuracy 100%) and MCI vs CN (accuracy 85%) and did not explore MCI conversion to AD. The ones that indeed evaluated MCI conversion like the longitudinal study from (Zhang et al., 2014) did not get such good results and showed high sensitivity results (83-100%) but poor specificity (46-88%). This was explained by the fact that positive results of PIB-PET were also present in other patients with other diseases (Lewy body dementia, Parkinson) and also in some normal subjects. For this reason, prior to ^{11}C-PIB-PET being widely used as diagnostic modality it is still important to demonstrate its accuracy, as it is a high cost investigation biomarker.

Furthermore, some AD studies were also based on cerebrospinal fluid (CSF) biomarkers. At present there are three main CSF possible biomarkers for AD molecular pathology: total tau protein (T-tau) that reflects the intensity of neuronal degeneration; hyperphosphorylated tau protein (P-tau) that probably reflects neurofibrillary tangle pathology; and the 42 amino-acid-long form of amyloid beta ($A\beta_{42}$) that is inversely correlated with $A\beta$ pathology in the brain (Cavedo et al., 2014). So, when using this modality for AD prediction these are the ones which are used as CSF features, like it was done in (Cheng et al., 2015).

In addition, genetic factors play an important role in late onset Alzheimer's disease (Albert et al., 2011). Variants of the apolipoprotein E (APOE) gene, found on chromosome 19, are known to affect the risk of developing AD. These are: APOE $\varepsilon 2$, which is relatively rare and may provide some protection against the disease; APOE $\varepsilon 3$, which is the most common allele, and is believed to play a neutral role in the disease (neither decreasing nor increasing risk); and APOE $\varepsilon 4$, which is the strongest known genetic risk factor for AD and present in about 25% to 30% of the population and in about 40% of all people with AD (Liu et al., 2013; Crenshaw et al., 2013). So people who develop Alzheimer's are more likely to have an APOE $\varepsilon 4$ allele than people who do not develop the disease.

Neuropsychological assessments have been used to grade the cognitive state of patients (Folstein et al., 1975) and to characterize dementia associated with AD in several studies

(Shankle et al., 1997; Salmon and Bondi, 2009; Chapman et al., 2010; Weintraub et al., 2012). A very common neuropsychological test is the MMSE already mentioned in this chapter when referring to (Shankle et al., 1997) study, one of the first using machine learning methods for AD diagnosis. The neuropsychological tests have proven to be extremely useful to identify the cognitive profiles, determine patterns of impairment, assess changes of impairment over time and also after treatment, and in fact, have been widely used clinically to achieve a probable diagnosis of the disease (McKhann et al., 2011). These are normally the preferred assessments used clinically because they present some advantages and facilities like being inexpensive in comparison with other types of exams, as neuroimaging, and are totally innocuous for the patients, compared to invasive tests like nuclear medicine imaging.

Almost all of the data from these different modalities described above, used as predictors for the disease, can be found in Alzheimer's Disease Neuroimaging Initiative (ADNI) data repository (http://adni.loni.usc.edu/data-samples/access-data/). All ADNI data is archived in a secured and encrypted system through Image Data Archive (IDA), of the University of Southern California's Laboratory of Neuro Imaging (LONI), and can be accessed with proper authorization provided by ADNI Data sharing and Publications Committee (DPC). This well-curated scientific data repository is remarkably successful across the globe, and has been a huge help for studies in this area providing data since 2004. According to (Murray, 2012) more than 1300 investigators have been granted access to ADNI data, resulting in extensive download activity that exceeds 1 million downloads of imaging, clinical, biomarker and genetic data. ADNI has data of AD, MCI patients and elderly cognitive normal (CN). These participants are followed and reassessed over time to track the pathology of the disease as it progresses.

2.1 Multimodality

Given that different modalities can help in the detection of different characteristics, an approach combining more than one modality would be preferable. Multimodal neuroimaging, which is the combination of more than one image modality, may play an important role with regard to early and reliable detection of subjects at risk of developing AD, for two important reasons. Firstly, neurodegeneration in AD cannot be reduced to a singular pathological process in the brain (Cavedo et al., 2014). Thus, different

modalities can detect different important neuropathological aspects which can facilitate the detection of the disease. Secondly, it is well accepted that the onset of appearance of these different neuropathological aspects in the brain may occur subsequently and not simultaneously (Cavedo et al., 2014). Therefore, depending on which stage the disease is, one modality may detect a certain pathological characteristic better than other. Intuitively, integration of more than one modality may uncover the previously hidden information that cannot be found using just a single modality. In this context, multimodality is seen as a very useful tool to achieve better classification performances (Zhang et al., 2012; Cavedo et al., 2014; Uludağ and Roebroeck, 2014). Several studies have exploited the fusion of the multiple modalities to improve AD or MCI classification performance. Some recent studies which use multimodality are presented in table 2.1. It is possible to see how using different modalities allows better results. For example, (Jie et al., 2015) tests two different multimodal cases MRI+PET and MRI+PET+CSF, using the same population, and shows that the second presents better results. In comparison with single modality, the multimodal approach requires a more careful handling of the data as in this case data from different modalities have to be joined and the number of features given to the classifier rises. The simplest way to combine the data, which was done in many studies (Bouwman et al., 2007; Vemuri et al., 2009; Walhovd et al., 2010) is by concatenating all the features of each modality in the same vector. Other more powerful methods include allocating kernels for each modality and then using a multi-kernel learning method to aggregate all kernels in one only kernel like it is done in (Hinrichs et al., 2011; Zhang et al., 2011).

Study	Subjects	Modalities	Classifier	AD vs CN	MCI vs CN	MCI-C vs MCI-NC
Zhang et al., 2011	51 AD + 99 MCI + 52CN	MRI + PET	SVM	90.6%	—	—
Zhang et al., 2011	51 AD + 99 MCI + 52CN	MRI + PET + CSF	SVM	93.2%	76.4%	—
Hinrichs et al., 2011	48 AD + 66 CN	MRI + PET	SVM	87.6%	—	—
Hinrichs et al., 2011	48 AD + 66 CN	MRI + PET + CSF+ APOE + Cognitive scores	SVM	92.4%	—	—
Huang et al., 2011	49 AD + 67 CN	MRI+PET	SVM	94.3%	—	—
Zhang et al., 2012	45 AD + 91 MCI + 50CN	MRI + PET + CSF	SVM	93.2%	83.2%	—
Gray et al., 2013	37 AD+ 75 MCI+ 35 CN	MRI+ PET+ CSF+ genetic	RF	89.0%	74.6%	58.0%
Liu et al., 2014	51 AD+ 99 MCI+ 52 CN	MRI+ PET	SVM	94.4%	78.8%	67.8%
Jie et al., 2015	51 AD+ 99 MCI+ 52 CN	MRI+ PET	SVM	95.0%	79.3%	68.9%
Jie et al., 2015	51 AD+ 99 MCI+ 52 CN	MRI +PET+ CSF	SVM	95.4%	83.0%	72.3%

Table 2.1: Accuracy of state-of-the-art Multimodality studies.

2.2 Feature Selection

As stated in the previous chapter, feature selection is a crucial step prior to any classification performed on high-dimensional data, because it reduces the number of features given to the classifier, leading therefore to sparse representations of data. As stated in (Yu, 2003), since 1970's the problem of feature selection has been extensively studied by the machine learning community and it has been proven that it can be effective to remove irrelevant and redundant features.

For neuroimaging studies there are essentially two ways of dealing with feature selection depending on the feature extraction chosen: 1- extracting specific brain regions, which are called regions of interest (ROI) and then performing feature selection in these regions; 2- using feature selection methods to select only the important features of the whole-brain. Many researchers opt for the first approach (Zhang et al., 2011; Zhu et al., 2015; Lahmiri and Boukadoum, 2013; Young et al., 2013) to avoid the high dimensionality problem. For example, in (Zhang et al., 2011) they select 93 ROIs and then average intensity of each ROI region having in total just 93 features for each neuroimage modality, in (Young et al., 2013) they choose to select 10 regions according to (Braak and Braak, 1995), therefore reducing drastically the computational time. However this approach can have some issues. First if there are no ROI's known a priori, and second having ROI's a priori will not give the possibility to find new important regions. On the one hand, it can be logical and facilitating to use just ROI's because they are based in previous studies which have proven a certain theory driven assumptions of which areas are most involved in the disease, nonetheless, research must be done not only to confirm findings but also to expand them, specially for AD case, where absolute knowledge is still on construction. Therefore, a whole-brain analysis may give new and interesting information, in addition to the one already known. On the other side, choosing to design a whole-brain classifier is quite challenging as it will present a lot more features. Typically the consequence is an overfitting of the data, leading to high accuracies for data used in designing the classifier, but poor classification accuracies for new independent test data. To overcome this problem, there is a huge need to use of feature selection methods.

A well-known method is the Least Absolute Shrinkage and Selection Operator (LASSO) method, introduced in 1996 (Tibshirani, 1996). This method uses a regularized L_1-penalty for sparse variable selection. The L_1-norm regularization was already used for

regression problems in several studies, but also has become popular topic in the context of classification problems. For the regression problems the most popular loss function used is the least squares estimate (Schmidt, 2005), for classification problems, however, least squares estimate may not be adequate, as it gives poor predictions compared to other loss functions, like logistic regression (Rosasco et al., 2004). The comparison of the L_1 penalty and the L_2 penalty was also performed, and results showed that models produced with L_1 penalty often outperformed the ones produced with L_2. (Schmidt, 2005).

Further work have also extended the LASSO for multi-task problems by using other regularization parameters, in particular the L_1/L_q regularization, with $1 \leq q \leq \infty$, proposed by (Yuan and Lin, 2006) for regression and extended for classification in (Meier et al., 2008). The L_1/L_q regularization belongs to the composite absolute penalties (CAP) family. When q=1, this extension is equivalent to the L_1-penalty problem; when q>1, the L_1/L_q regularization promotes group sparsity in the resulting model, which is quite desirable in many applications of classification problems. Studies analyzed also the influence of different values for q. For example, the results of (Liu and Ye, 2010) showed that smaller values of q had lower balanced error rates then higher values of q. Another study also performed this analysis of the q values, in this case for a multi-task learning with large scale experiments (Vogt and Roth, 2012), and the results also showed better results for lower values of q more precisely for values between 1.5 and 2. Thus, L_1/L_2 regularization seem to be the preferred choice to use for the model creation. This L_1/L_2 norm penalty was already used for AD studies. In (Zhou et al., 2011) they used it for regression, specifically they tested a multi-task regression problem predicting disease progression measured by cognitive tests. This method of using LASSO with the L_1/L_q penalty for multi-task is also known as the the group LASSO multi-task method or just group LASSO, for simplification.

2.3 Classifiers

Most of machine learning studies carried out for AD classification from neuroimaging problems used the Support Vector Machine (SVM) algorithm as the classifier (Zhang et al., 2011; Hinrichs et al., 2011; Liu et al., 2014; Jie et al., 2015), due to its good accuracy, ability to cope with very high-dimensional data (several features and small

number of examples) as it uses kernels, and for providing a unic solution every time the problem is solved with the same inputs and same conditions. In fact, all studies presented in table 2.1 used SVM as the classifier, expect for (Gray et al., 2013) which used random forest (RF). Although SVM can bring good results and has big advantages, this classifier is a non-probabilistic binary classifier which uses a supervised learning algorithm, and the use of other classifiers, like probabilistic classifiers, could also be helpful for these kind of studies.

One example is Logistic Regression (LR) which is able to predict, given a sample input, a probability distribution over a set of classes, rather than only outputting the most likely class that the sample should belong to. So, it provides classification with a degree of certainty. Nonetheless, it is worth noting that LR should not be used when there are a large number of features in comparison to the number of training samples, because of the problem of overfitting. In this sense, sparse models are necessarily needed for neuroimaging pattern recognition studies, which is achieved by adding regularization parameters as described above in section 2.2. In (Rao et al., 2011) they tested LR with L_1 and also another penalty, the L_2, for classification of AD vs CN based on sMRI, and presented classifications with better accuracies when using L_1 than when using only the L_2 regularization. Furthermore, (Ryali et al., 2010) used LR with a combination of L_1 and L_2 norm regularization for whole-brain classification of fMRI data, however not for AD applications. With their work they could identify relevant discriminative brain regions and accurately classify fMRI data.

Study	Subjects (MCI-NC; MCI-C)	Modalities	Conversion period	Accuracy	AUC
Nho et al., 2010	355 (205; 150)	MRI + APOE+ family history	0-36 months	71.6%	—
Zhang et al., 2011	99 (56, 43)	MRI + PET + CSF	0-18 months	(sens 91.5% spec 73.4%)	—
Davatzikos et al., 2011	239 (170, 69)	MRI + CSF	0-36 months	61.7%	—
Hinrichs et al., 2011	119	MRI + PET + CSF+ APOE +Cognitive scores	0-36 months	—	0.7911
Ye et al., 2012	319 (177, 142)	MRI+ APOE+ cognitive scores	0-48 months	—	0.859
Young et al., 2013	143 (96, 47)	MRI+ PET +APOE	0-36 months	74.1%	0.795
Cheng et al., 2015	99 (56, 43)	MRI + PET + CSF	0-24 months	80.1%	0.852

Table 2.2: Performance of different state-of-the-art Multimodality studies for predicting MCI conversion to AD.

Recently two papers have also explored another classification algorithm for AD studies: the Gaussian Process (GP). Gaussian Process is a probabilistic classifier and can be seen as a Kernelised Bayesian extension of logistic regression. The first study to use GP for AD applications was (Young et al., 2013). They tested GP in a multimodality

study using MRI, PET, CSF and genetic for classification of MCI-C vs MCI-NC. Their multimodality results were significantly better than any single modality and better than when performing the classification with SVM. Thus, they achieved state-of-the art accuracy (table 2.2) and showed that the GP classifier can be successfully applied to the prediction of conversion of MCI patients to AD. Another study (Challis et al., 2015) also used GP classifier, but only for fMRI data and tested different covariance functions. These papers showed that with GP the results were very similar to SVM results or even better. Additionally, they also argue that GP has advantages in comparison to SVM: (i) having a probabilistic classification means that each diagnosis includes an attached degree of confidence rather than a simple binary decision. In case of clinically decision this can be quite useful, as frequently this decision is hampered by overconfidence. (ii) GP is better at finding a set of kernel weights for optimum classification. Rather than finding these through a grid search, like is done when using SVM, with GP a tuning is performed via the likelihood function, which seems to be both more robust and allows a wider search range.

From table 2.1 it is possible to note that the older studies were more concerned with classifying AD vs CN and CN vs MCI. However more recent studies are also concerned in classifying MCI-C vs MCI-NC. Some of the studies performed to predict MCI conversion to AD are presented in table 2.2.

Chapter 3

Proposed Methods

In the following chapter the proposed methods will be explained, more precisely each section will present a brief introduction about the basic theory behind each method and the toolboxes used to implement them.

3.1 Feature Selection

The overall goal of feature selection is to overcome the high-dimensional problem by selecting the most important features, i.e., the ones that help minimizing redundancy and maximizing relevance in distinguishing particular conditions of interest. In the case of this work, to distinguish two classes from the study (binary or binomial classification). In this perspective, the goal is to get sparse models, i.e., models that will have zero weights for the features that are not relevant in distinguishing the classes.

3.1.1 LASSO with L_1 penalty

Several studies that wish to achieve sparse models, are using the L_1 norm penalty as it has a strong sparsity-inducing property and has shown great empirical success for various applications (Schmidt, 2005; Liu and Ye, 2010). The least absolute shrinkage and selection operator (LASSO) is an example of a model trained with the L_1-norm regularization.

Given a dataset D with a feature data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ having n samples and d features, one can represent the dataset by D=$\{\mathbf{x}_i, y_i\}_{i=1}^{n}$, where $\mathbf{x}_i$ is the feature vector representing the i-th sample which corresponds to the i-th row of matrix $\mathbf{X}$ and $y_i \in \{-1, 1\}$ is the respective label for that i-th sample. Mathematically, the objective function to minimize in order to determine the features weight vector $\mathbf{w} \in \mathbb{R}^{d \times 1}$ is presented in equation 3.1, where $f(\mathbf{w})$ is the loss function and constant λ is the parameter that will determine the contribution of the L_1-norm of the weight vector $\mathbf{w}$. The loss function can be the least-squares loss and so this equation can be written as 3.2, or in case of choosing the logistic loss function this equation would be represented by 3.3.

$$\min_{\mathbf{w}} f(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 \tag{3.1}$$

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1 \tag{3.2}$$

$$\min_{\mathbf{w}} \sum_{i=1}^{n} s_i \log(1 + exp(-y_i(\mathbf{w}^T \mathbf{x_i} + b))) + \lambda \|\mathbf{w}\|_1 \tag{3.3}$$

In equation 3.3 the value s_i is the weight for the i-th sample and b represents the intercept (scalar) also seen as the bias.

As it was mentioned in chapter 2 the logistic loss function was already tested when studying classification problems showing better performance than the least-squares (Rosasco et al., 2004), therefore, instead of choosing the least squares as the loss function, this work will be focused on using the logistic loss represented by equation 3.3.

3.1.2 Group LASSO multi-task with L_1/L_2 penalty

LASSO was further extended to many other variants. These were created to make the method more useful also for other particular problems, for example when having a multi-task problem. Thus, the L_1/L_q regularization ($\lambda \|\mathbf{w}\|_{q,1}$) emerged, also known as group LASSO multi-task. As already stated in chapter 2, (Liu and Ye, 2010; Vogt and Roth, 2012) tested different values for q and showed better results for lower values of q more precisely for values between 1.5 and 2. Hence, L_1/L_2 regularization will be used for the model creation. Considering this regularization, minimizing the objective function, in a multi-task problem, can be performed by the following equation:

$$\min_{\mathbf{w}} \sum_{j=1}^{t} \sum_{i=1}^{n} s_i \log(1 + exp(-y_{ij}(\mathbf{w}^T \mathbf{x_{ij}} + b_j))) + \lambda \|\mathbf{w}\|_{2,1} \tag{3.4}$$

where t denotes the number of tasks. It is easily seen, by comparing equations 3.3 and 3.4, that this group LASSO multi-task problem reduces to the LASSO method when there is only one task (t=1) and for q=1. The difference in this method is that for Group LASSO multi-task with L_1/L_2 penalty the algorithm performs the sum (i.e. the L_1 norms) of the L_2 norms of the weight vector $\mathbf{w}$ for each feature over all tasks. In other words the L_2 norm of each weight feature vector will form a group and L_1 norm will select the features in accordance to the weight of each group formed. Consequently, it tends to select features based on the strength of each feature over all t tasks (Zhou et al., 2011).

When analysing a multimodal problem this technique can also be used, and in this case, each modality will be seen as a single task. A illustration comparing these two feature selection methods for a multimodal problem is presented in figure 3.1. The figure shows in A the feature selection with L_1-norm, here the feature selection is performed independently on each modality, and in B the L_1/L_2 penalty, which demonstrates that a common set of features are selected based on information of both modalities.

The toolbox Sparse Learning with Efficient Projections (SLEP) (Liu et al., 2009.), implemented in MATLAB, is widely used to perform these and many other regularization's and has been shown very effective on many datasets. Plus, in terms of computational time, it is quite good at handling large-scale data in order to create sparse data. The first version of the SLEP toolbox was released in August 2009, currently its latest version is Version 4.1., released in December 2011.

For both LASSO and Group LASSO, SLEP toolbox automatically computes a $\lambda_{\max}$ (the maximum value of λ) and gives the user the possibility of establishing values in the interval [0,1] for the percentage of this $\lambda_{\max}$. Thus, the resulting regularization parameter used in the program is given by multiplying the maximum value $\lambda_{\max}$ by a chosen ratio of the regularization parameter, which will be presented as λ, in this thesis.

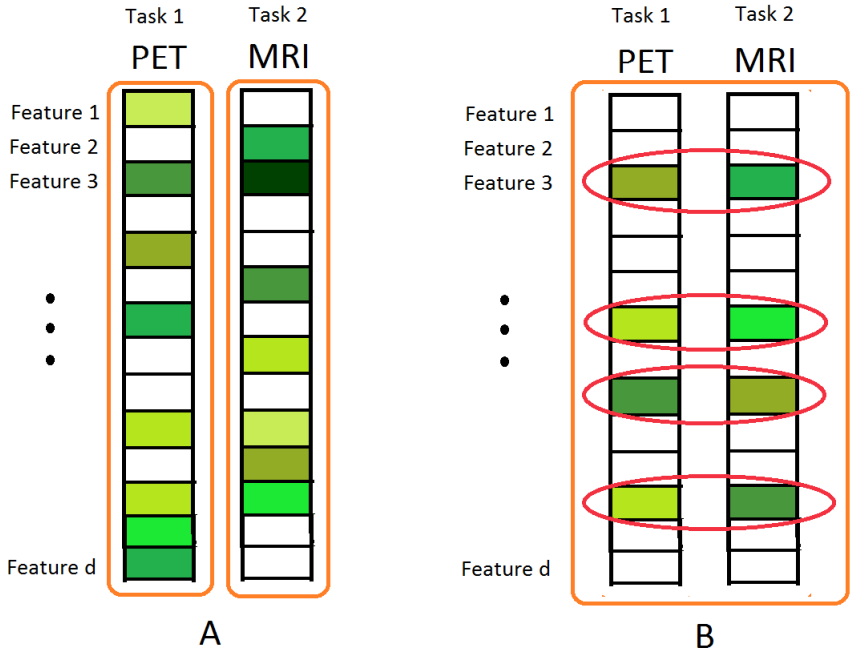


Figure 3.1: Feature selection for a multimodal (PET + MRI) case. A - Using the L_1 norm as the regularization. B - Using L_1/L_2 norm as regularization. Adapted from (Liu et al., 2014)

3.2 Classifiers

In this section the proposed classifiers will be discussed and compared. All of these classifiers have the advantage of returning the same output upon feeding the classifier with the same input.

3.2.1 SVM

The support vector machine (SVM) is a non-probabilistic decision machine classifier and so it does not provide direct posterior probabilities, in other words it does not give the probability of belonging to a certain class taking into account previous examples, instead it gives directly the class of a given example which is attempting to classify. This classifier is extremely useful in neuroimaging studies as it deals well with high-dimensional data by using the kernel method, as stated previously in section 1.3. Furthermore, it also has the advantage of giving a unique best solution for the problem of classification.

Likewise defined previously, assuming we have a dataset of n examples D=$\{\mathbf{x}_i, y_i\}_{i=1}^n$, consisting of features vectors $\mathbf{x}_i$ with dimensionality d (number of features in each feature vector $\mathbf{x}_i$) and the respective correct label $y_i \in \{-1, 1\}$, ; the objective is to define a function based on the data that will accurately predict the labels y_i for new feature vectors $\mathbf{x}$, i.e. $f(\mathbf{x}) = y$. For this purpose, SVM uses a linear model for binary classification which has the form:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \qquad \text{Classification: } y(\mathbf{x}) = \text{sgn}(\mathbf{w}^T\mathbf{x} + b) \tag{3.5}$$

where b is the bias term and $\mathbf{w}$ is the weight vector which is a normal vector perpendicular to the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$. The new data points $\mathbf{x}$ will be classified according to the sign of $\mathbf{w}^T\mathbf{x} + b$. Thus, if $\mathbf{w}^T\mathbf{x} + b \geq 1$ then $y = 1$; if $\mathbf{w}^T\mathbf{x} + b \leq -1$ then $y = -1$; these constraints can be summarised by: $y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0$.

In case $d = 2$, i.e. when having just two features for each sample, this function will draw a line on a graph of x_1 vs x_2 separating the two classes, see figure 3.2. For other cases ($d \geq 2$), which have the most prevalence, SVM will define a hyperplane on graphs of $x_1, x_2, \ldots, x_d$ (Fletcher, 2006). This hyperplane represents the decision boundary, i.e., all points lying on one side of the hyperplane and with sgn($\mathbf{w}^T\mathbf{x} + b$) positive will be classified as having $y = 1$ and the points lying on the other side with sgn($\mathbf{w}^T\mathbf{x} + b$) negative will be considered $y = -1$. SVM defines this decision boundary by using a subset of data points known as support vectors (examples closest to the separating hyperplane). Among all hyperplanes separating the data, there exists one which is optimal, this hyperplane would be the one representing the largest separation, or margin, between the two classes. The margin is defined as the perpendicular distance between the decision boundary and the support vectors and represented by $\|\mathbf{w}\|^{-1}$ see figure 3.2 for illustration. Therefore SVM needs to find values of $\mathbf{w}$ and b that will maximize $\|\mathbf{w}\|^{-1}$ which is equivalent to minimizing $\|\mathbf{w}\|^2$ or minimizing $\frac{1}{2}\|\mathbf{w}\|^2$ (Bishop, 2006).

This is an example of a quadratic programming optimization problem. We need to minimize a quadratic function taking into account a set of linear inequality constraints:

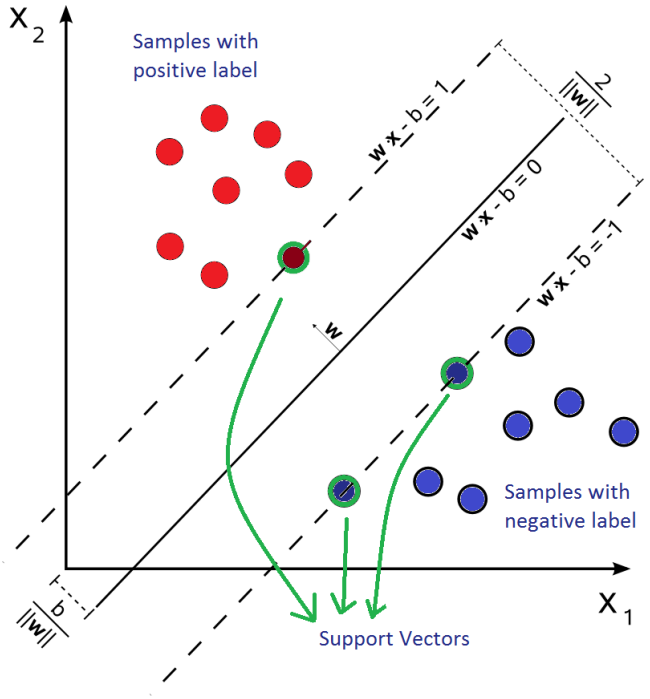


Figure 3.2: A 2D representation of the hyperplane $\mathbf{x}_i.\mathbf{w} + b = 0$ defined by the support vectors (shown in green and lying on the margins) and maximum-margin $\|\mathbf{w}\|^{-1}$ to separate class positive from class negative. Adapted from (https://en.wikipedia.org/wiki/Support_vector_machine).

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \qquad s.t. \qquad y_i(\mathbf{x}_i.\mathbf{w} + b) - 1 \geq 0 \qquad \forall i \tag{3.6}$$

To solve this constrained optimization problem we will need to allocate for each constraint in 3.6 Lagrange multipliers $\alpha_i \geq 0$ (Fletcher, 2006).

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \alpha[y_i(\mathbf{x}_i.\mathbf{w} + b) - 1 \qquad \forall i] \tag{3.7}$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i[y_i(\mathbf{x}_i.\mathbf{w} + b) - 1] \tag{3.8}$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \alpha_i y_i(\mathbf{x}_i.\mathbf{w} + b) + \sum_{i=1}^{n} \alpha_i \tag{3.9}$$

The Lagrangian L has to be minimized with respect to the variables $\mathbf{w}$ and b and maximized with respect to the variables α_i (Scholkopf and Smola, 2002). To do this we have to compute the derivatives of L with respect to $\mathbf{w}$ and b and setting the derivatives

to zero:

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{3.10}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.11}$$

However, in practice defining a hyperplane can be inappropriate in cases which data is not fully linearly separable. For example, when the data has some misclassified examples, i.e., some training samples have incorrect label and are on the wrong side of the decision boundary, defining a linear decision boundary will not be possible. So, in order to allow some misclassified examples the introduction of slack variables ($\xi_n \geq 0$) with one slack variable for each training data point is needed. The way to get this trade-off between maximizing the margin and minimizing the number of misclassified sample is to introduce a parameter C and minimize the function:

$$C \sum_{i=1}^{n} \xi_n + \frac{1}{2} \|\mathbf{w}\|^2 \tag{3.12}$$

In the limit when C$\rightarrow \infty$ this model will not allow soft margins and therefore it will recover the previous SVM model for linearly separable data.

Furthermore, given the duality property of Lagrange multipliers, this problem can be rewritten in dual form, which corresponds to writing the algorithm from equation 3.5 in terms of the inner product between points in the input space giving rise to equation 3.13, taking into account equation 3.10. The fact that it is possible to express the algorithm in its dual form will be very useful for later application of the kernel trick, which will be explained in the following section 3.2.1.1.

$$f(\mathbf{x}) = \text{sgn}(\sum_{i=1}^{n} \alpha_i y_i (\mathbf{x} \cdot \mathbf{x}_i) + b) \tag{3.13}$$

3.2.1.1 Kernel Trick

Another way of dealing with data not linearly separable in the input space, is to map the data into a higher dimensional space, named feature space, and perform the linear separation for example using a kernel in this new space. As already stated, having the problem represented in its dual form enables the use of kernels. There are many kernel functions available, the simplest kernel and one which is often used in these problems is the linear kernel, which is basically the use of the dot product. One of the big advantages of using kernels in SVM is that it overcomes the high-dimensional problem and therefore makes the problem computationally less demanding.

Given two feature vectors, kernel linear function will perform the dot product of these two feature vectors and will return a real number characterizing their similarity. Therefore, instead of having a data matrix with dimensionality of number of features per number of samples ($d \times n$) this kernel transformation turns the data matrix into a much smaller dimension: number of samples per number of samples ($n \times n$).

3.2.1.2 Nested Cross-validation to discover C parameter

Normally, for pattern recognition techniques in neuroimaging data, the number of samples available are not very high, introducing some issues because in a predictive test having many examples is crucial for the model creation and further evaluation. In this sense, usually the technique of choice to estimate the performance of the predictive model is the cross-validation method. A common variant of cross-validation is called "leave-one-out" and consists in three main steps which should be followed: leaves one example out and trains with the remaining ones to make a prediction for this example; repeats this for every example in turn (Pereira et al., 2009); and then compares with the actual values and the statistics of the predictions made for each example are calculated. Finally all of the results of the statistics are averaged and these final results will represent the evaluation of the model created. This approach was also extended for the k-folds. In k-folds cross validation the dataset is partitioned into k different test and training sets and then the statistics is averaged over the k-folds.

In case of the SVM classifier, besides evaluating the performance of the classifier, an evaluation to determine which parameter C is the best for the studying data is also

needed. Therefore, the usual procedure is to use a Nested Cross-validation. In other words, it performs a cross-validation inside the previous cross-validation. Thus, the dataset is first divided into the k different training and testing sets (outer loop), then each of these created training sets is further divided into other training and testing sets (inner loop) which will test different values of C in a grid search and after determining which C parameters performed better the optimal C will then be used for the outer loop. An example of the nested cross-validation is presented in figure 3.3 with 5-Folds for the outer loop and 2-Folds for the inner loop.

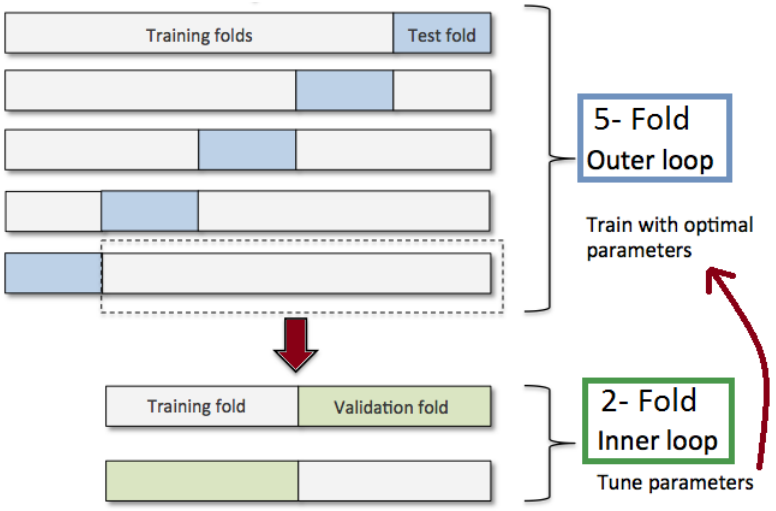


Figure 3.3: Diagram representing the nested cross-validation, in this case the outer loop has 5-folds and the inner loop has 2-fodls. The optimal parameters are obtained in the inner loop and further given to the outer loop and then the evaluation of the model is performed in this outer loop (http://sebastianraschka.com/faq/docs/evaluate-a-model.html).

To implement the SVM one can use the Library for Support Vector Machines (LIBSVM) tool (Chang and Lin, 2011) created by Chih-Chung Chang and Chih-Jen Lin from Department of Computer Science, National Taiwan University. The initial version of this tool was released in 2000 and currently is on its Version 3.21, released on December 14, 2015. Since its release, LIBSVM has been widely used in many areas such as Neuroimging and Bioinformatics studies, counting for more than 250,000 downloads of the package in the period from 2000 to 2010.

3.2.2 Logistic Regression

In contrast with SVM, Logistic Regression (LR) is a probabilistic classifier, so the result of the classification is not the sign which will represent the class label like in SVM, instead the result will be a probability.

LR is based on the concept of linear regression models, however in this case it has to be generalised in order to be used for classification. In linear regression models the model prediction is given by the linear function: $y(x) = \mathbf{w}^T\mathbf{x} + b$, with $y \in \mathbb{R}$, consequently it would have values ranging [-∞, + ∞]. However, in classification problems, we wish to predict discrete class labels, or in this case, a posterior probability which will identify how likely a given sample belongs to the positive or negative class. Therefore the values of y should be in the interval [0,1], as it corresponds to probability values. In order to achieve this, the linear function has to suffer some transformation, which is done with a non-linear function called logistic sigmoid function $\sigma(.)$, represented by equation $\sigma(a) = \frac{1}{1+exp(-a)}$. Thus, the probabilities in logistic regression for a given sample belonging to a positive class are given by equation 3.14. The graphical representation of the sigmoid function is presented in figure 3.4.

$$p(\mathbf{w}^T\mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}} \tag{3.14}$$

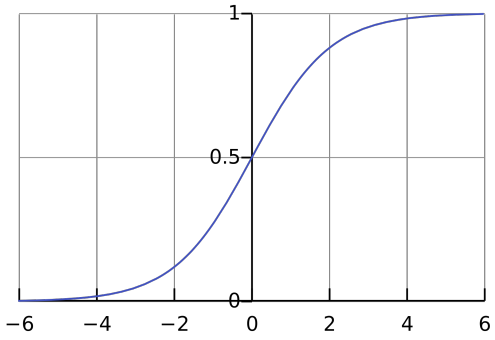


Figure 3.4: Representation of the logistic sigmoid function σ(a). Regardless the value of a (horizontal axis), the function σ(a) will always return a value in the interval [0,1].

Plus, this sigmoid function satisfies the following symmetric property: $\sigma(-a) = 1 - \sigma(a)$; which allows finding the corresponding probability of the same sample belonging to a

negative class and is in accordance with the probability property that P(positive class) + P(negative class) = 1.

The implementation of this classifier is quite straightforward when using the SLEP toolbox that was previously used for feature selection. First the SLEP toolbox is used to determine the weight vector **w** and the parameter b, then the probabilities are discovered by using equation 3.14.

As LR returns probabilities, in order to turn these into a discrete label to further use statistics to evaluate the model prediction performance, a threshold should be defined. Normally this threshold can be defined as 0.5, meaning that all the test samples with probabilities inferior to 0.5 are classified as negative and all the ones which are higher than 0.5 are classified as positive label, the ones with values equal to 0.5 have equal probability of belonging to positive class and negative class. This means that the classifier cannot decide which class it should belong to. In these cases, it is possible to predefine the location of a equal sign in the threshold, which will determinate the class to attribute for that sample. For example, if one defines that only values higher than 0.5 are seen as positive samples, it implies that values equal or lower will be seen as negative. In this case, the decision is not made by the classifier instead is the user who defines it.

3.2.3 Gaussian Process

Gaussian process classifiers are formulated based on the foundation of linear logistic regression models, more specifically they can be seen as a kernelized Bayesian extension of the logistic regression model (Young et al., 2013).

3.2.3.1 Bayesian theory

Before starting to explain GP classifier theory, it is useful to understand how the Bayesian theory is formulated. This theory is known as:

$$p(h|D) = \frac{p(D|h)p(h)}{p(D)} \quad , \quad p(D) \neq 0 \tag{3.15}$$

Interpreting equation 3.15 we can say that $p(h \mid D)$ is the posterior probability, i.e., is the probability of a hypothesis h after data D is observed; $p(h)$, is the prior probability, which is the probability of h before any D is observed; and $p(D \mid h)$ also known as the likelihood is the probability of D with h fixed.

For most standard Bayesian methods, the prior distribution $p(h)$ is fixed before any data is observed. However, for GP classifiers as it is a supervised classifier, we want to determine the statistical inference based on a prior distribution which is estimated from the data, i.e., empirically. For that reason, in GP classifiers the method used is called Empirical Bayes method, which is an approximation of the fully Bayesian treatment.

3.2.3.2 Gaussian Process classification formulation

The probability of belonging to a certain class at an input location has a relation with the value of some latent function f at that location (Young et al., 2013). For the purpose of determining this probability, GP classifier forms a prior over this latent function $p(f)$, and then squishes this through the logistic sigmoid function $\sigma(.)$, as in logistic regression, to obtain values in the [0,1] interval, guaranteeing valid values for probabilistic interpretations.

Thus, given a dataset $D = (x_i, y_i)_{i=1}^{n}$ with n samples, to obtain the posterior probabilities mathematically these can be represented by (Williams, 2006):

$$p(f|D) \propto p(f)p(D|f) \tag{3.16}$$

$$\text{posterior} \propto \text{prior} \times \text{likelihood} \tag{3.17}$$

In order to perform GP classification and apply a prior over the function values, the classifier should be first parametrized by a mean function and a covariance function (equation 3.18).

$$p(f) = \mathcal{N}(mean, cov) \tag{3.18}$$

The covariance function plays an important role in GP classifier formulation because it will define the similarity between data points (samples). Usually, this is achieved using the linear kernel as the covariance function k, which performs the dot product between two samples $\mathbf{x}$ and $\mathbf{x}'$ and is parametrized by a hyper-parameter θ giving rise to a covariance matrix K, see equation 3.19. Thus, equation 3.18 can be rewritten as equation 3.20. This covariance function can then be optimized by determining the proper hyper-parameter θ via type-II maximum likelihood, also known as Empirical Bayes. Type-II maximum likelihood maximizes the marginal likelihood which will automatically incorporates a trade-off between model fit and model complexity (Williams, 2006).

$$K = k(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}' + \theta \tag{3.19}$$

$$p(f|\theta) = \mathcal{N}(mean, K) \tag{3.20}$$

Once the hyper-parameters are defined, the predictions to unseen data are done by integrating across this prior (Young et al., 2013). However, for classification problems, because the sigmoid function is used, the exact Bayesian inference is analytically intractable and therefore approximation inference techniques have to be used, in order to obtain the final desired posterior probability $p(f|D, \theta)$. For this purpose, the usual approximation methods used are the Laplace Approximation (Williams and Barber, 1998) or the Expectation-Propagation (EP) method (Minka, 2001).

Likewise LR, the GP will also need to define a threshold in order to further transform these probabilities into class labels for further statistical evaluation.

To implement the GP classifier the Gaussian Processes for Machine Learning (GPML) Matlab Code version 3.6 was used (Rasmussen and Nickisch, 2015). The code was written by Carl Edward Rasmussen and Hannes Nickisch and is based on previous versions written by Carl Edward Rasmussen and Chris Williams. Both Carl Edward Rasmussen and Chris Williams are the authors of the book Gaussian Processes for Machine Learning (Rasmussen and Williams, 2006) which explains in detail how GP are formulated and used for pattern recognition problems. This code runs on both Octave 3.2.x and Matlab 7.x and later.

Chapter 4

Methodology and Results

This chapter will present the experiments performed in this thesis and its results along with the discussion. The first section 4.1 shows the details of the data used, section 4.2 presents the experimental approach tested in this work, finally in section 4.3 the results and discussion are presented.

4.1 Data

4.1.1 Subjects

All data used in this thesis was collected from the ADNI database and were baseline scans, i.e. scans corresponding to the subject's first visit. The total number of subjects was 210, with 50 belonging to the AD group, 48 to the CN group and 112 to MCI, which in turn, were further divided into 82 MCI-non converters and 30 MCI-converters, over a 24 month period, by analysing an excel file provided by the ADNI database where the diagnosis history is presented. The demographic data of these subjects is presented in Table 4.1.

Prior to classification, some statistical tests on these demographic data are necessary to make sure that the classification results for the different groups are not influenced by a significant difference in age populations or by a disproportion of males or females subjects. Hence, two statistic tests were performed: two sample t-test for age and the Chi-squared test for homogeneity of gender.

	AD (50)	**CN (48)**	**MCI (112)**	**MCI-C (30)**	**MCI-NC (82)**
Male % (subjects)	56% (28)	62.5% (30)	61.61% (69)	56.67% (17)	63.4% (52)
Female % (subjects)	44% (22)	34.8% (18)	38.39% (43)	43.33% (13)	36.6% (30)
Mean age ± std	76 ± 7	76 ± 5	75 ± 7	74 ± 8	75 ± 7

Table 4.1: Distribution of male and female subjects, with total number of subjects presented in parenthesis, and mean ages with the respective standard deviation values for each group.

4.1.1.1 Two sample t-test for age

The two-sample t-test is a test statistic which follows the t-Student's distribution under a null hypothesis which states that the means of two independent samples are equal. If the p-value obtained from this test is lower than a significant level previously defined (usually 0.05 or 0.1) then the null hypothesis should be rejected. In order to determine if the various groups of data (AD; MCI; CN; MCI-C; MCI-NC) are significantly different from each other, in terms of their mean ages (μ), the following the null hypothesis were formulated:

Null hypothesis:

$H_0 : \mu_{AD} = \mu_{CN} \quad H_0 : \mu_{AD} = \mu_{MCI} \quad H_0 : \mu_{CN} = \mu_{MCI} \quad H_0 : \mu_{MCI_C} = \mu_{MCI_{NC}}$

The p-values for these four tests were respectively: 0.9536, 0.1585, 0.1513 and 0.7222. Thus, for all the t-tests performed the p-value obtained was higher than the level of significance predefined (0.05), which implies that the null hypothesis is not rejected and consequently the mean ages of the different groups are not significantly different.

4.1.1.2 Chi-squared test for homogeneity for gender

In the case of statistical test for gender, t-test is not adequate because the variable is not continuous. In this regard, rather than using the t-test, a statistical test for categorical variables should be used. A good example is the Chi-squared test for homogeneity, in this case, to test homogeneity of gender in each group. For this test the hypothesis formulated were the following ones:

Null hypothesis:

H_0 : Number of males from AD = number of females from AD

H_0 : Number of males from CN = number of females from CN

H_0 : Number of males from MCI = number of females from MCI

H_0 : Number of males from MCI-C = number of females from MCI-C

H_0 : Number of males from MCI-NC = number of females from MCI-NC

Alternative hypothesis:

H_1 : At least one of the null hypothesis statements is false

The p-value obtained was 0.9064, which is greater than the level of significance predefined (0.05). Therefore the null hypothesis will not be rejected, which means that the Chi-squared test performed did not find significant differences in gender proportions for each group from this study.

4.1.2 Images

4.1.2.1 MRI images

The MRI brain images used from these subjects were acquired on 1.5T MRI scanners using volumetric T1-weighted sequences to map brain structures. The images had been preprocessed earlier by (Morgado, 2014) previous work. The preprocessing they went through included skull-stripping, i.e. the brain tissue in all MR images was extracted, and tissue segmentation, where the brain tissue was segmented into white-matter (WM) and gray-matter (GM), producing probability maps of GM and WM for each MR image. The brain extraction was done using the FreeSurfer, which is an open source software suite for processing and analyzing brain MRI images. While the segmentation was done with Statistical Parametric Mapping (SPM) version 8. SPM is a widely used software package which was designed for the voxel-based analysis of brain imaging data. Using the DARTEL toolbox from SPM8 an iterative non-linear registration into a subject-specific template was performed. Further details of this preprocessing can be found in (Morgado, 2014). All images were then warped into the Montreal Neurological Institute (MNI) 152 standard space, and therefore the final images were in a coordinate space of MNI 152 template with $1.5 \times 1.5 \times 1.5$ mm^3 resolution and 121×145×121 (in x y z) matrix dimension giving a total of 2122945 features for each MRI image.

4.1.2.2 PET images

The FDG-PET images collected from the ADNI database also did not need any further processing as they were already preprocessed by the Banner Alzheimer's Institute (Arizona) of the ADNI PET Core using the SPM5 and then uploaded to the LONI website. The essential preprocessing step they when trough was the spatial normalization. Each image was re-centered to correspond to the center of the SPM MNI template space. This spatial normalization will adapt the different shapes of the brain scans so one location in one subject's brain scan corresponds to the same location in another subject's brain scan, and thus facilitate further voxel-to-voxel analysis. The resulting images were in the coordinate space of the SPM template with 2 cubic mm voxel size and $79 \times 95 \times 68$ (in x y z) matrix dimension giving a total of 510340 features.

Although both modalities used MNI templates, they are registered in different spaces resulting in different matrix dimensions and consequently different total number of voxels for the whole-brain. In order to use the group LASSO multi-task feature selection from SLEP toolbox, a necessary condition is having the same spatial space for both modalities and having the same number of features. This requires further data processing, to transform both data images into the same spatial space. For this transformation, all PET images were reshaped to have same matrix dimensions as MRI images.

4.2 Experimental Design

This thesis explores the combination of data from two modalities, PET and MRI, using the group LASSO feature selection with L_1/L_2 penalty and three different classification algorithms (LR, GP, SVM). The objective is to compare the results between these different classifiers and analyze the relation between the lambda value (λ) from equation 3.4 and the accuracy of each classification procedure and finally to analyze the advantage of using probabilistic classifiers. The diagram of the experimental design used for this multimodal approach is presented in figure 4.1. For the single modal test a similar scheme was followed by using just LASSO with L_1 penalty from equation 3.3.

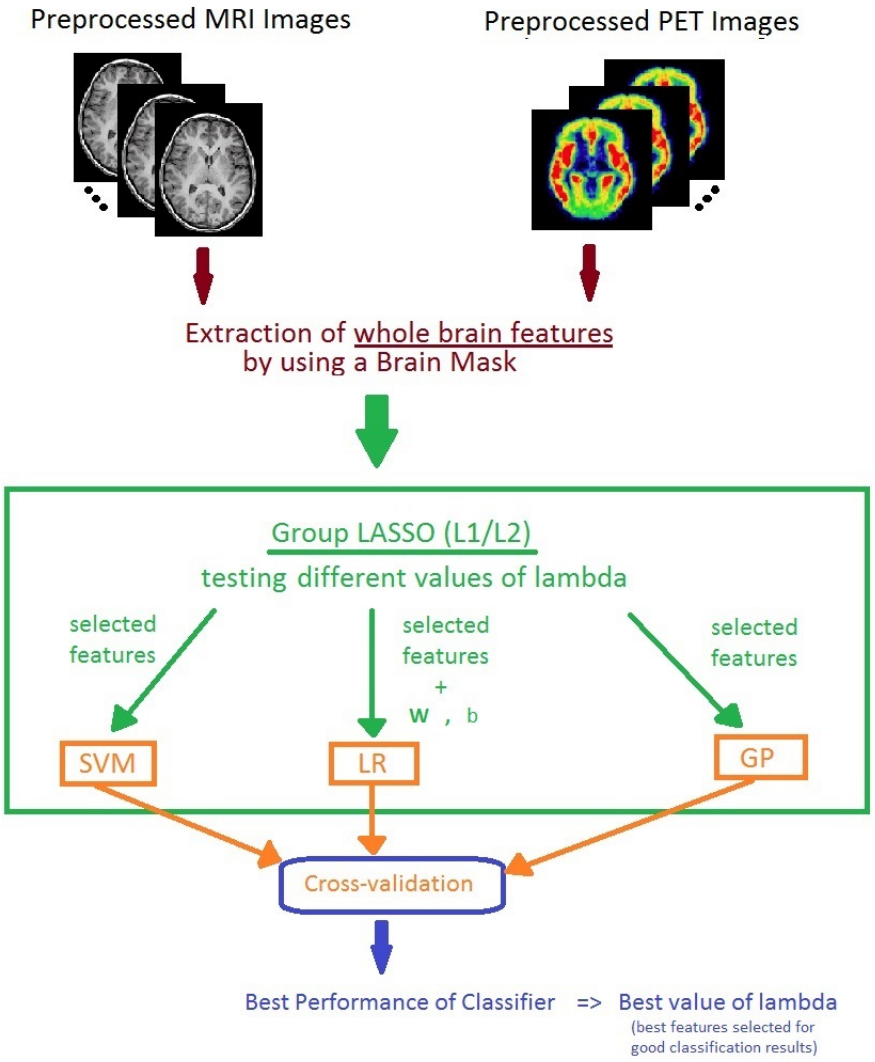


Figure 4.1: Representation of the framework for the multimodal classification problems.

4.2.1 Feature Extraction

One of the first steps performed to handle the high-dimensional problem when dealing with neuroimages, in a whole-brain approach, is extracting the voxel intensity features only from the brain. This is done because voxels outside the brain evidently won't contribute to distinguish the different subjects and would only lead to an unnecessary computational cost. Therefore, this approach of selecting only brain voxels, will help reduce the number of features given to the classifier and can be easily accomplished by selecting only the voxels from a brain mask previously defined based on the images. In case of MRI data, all images were registered to the MNI 152 space and the corresponding brain mask created for this space was available, consequently the selection of features

was straightforward. For PET data, however, there was no predefined brain mask for the images downloaded from ADNI database and thus this mask had to be created. The mask was created by averaging the PET images from all subjects (AD, CN and MCI) and then converting the image from a grayscale image to a binary image by thresholding the averaged image in order to get only voxels from the whole-brain with value 1 and the rest with value 0; if any 0 values were attributed inside the whole-brain these were converted to 1. The threshold value was 0.5, this value was determined empirically so the resulting mask could adjust to the brain shape from the averaged image obtained. By performing this feature elimination step it was possible to reduce the data matrix dimension by 26% for MRI images, having now 557780 features, and by 50% for PET images, having now 256627 features. Recall, however, that PET images had to be transformed into the same space as MRI images, thus, at the end of this processing step and prior to feature selection, both modalities had 557780 features for the whole-brain, which were provided to the classifier.

4.2.2 Feature Selection

After feature extraction, the total number of brain voxels (557780 features) is still quite high. Therefore, a feature selection method should be used in order to reduce the costly computation and avoid overfitting. In this work, the feature selection step was done by using the SLEP toolbox, more specifically by computing the LASSO (L_1 regularization) for the single modality approach and for the multimodal case by using the Group LASSO (L_1/L_2 regularization), both with logistic loss function, as described previously in section 3.1.

This method returns the weight vector $\mathbf{w}$ with size 557780 for the single modality case or a $\mathbf{W}$ matrix with size 557780×2 (number of features $\times$ number of modalities) for the multimodal case. Each line i of this matrix shows the weight given for a determinate feature i for the different modalities (first column for PET, second column for MRI). By analyzing each line from the matrix $\mathbf{W}$ it is possible to locate the selected features. All the lines with values different from 0, i.e. all the features with weights different from 0 are considered as the selected features. The level of sparseness provided by the algorithm will depend on the parameter λ. In SLEP toolbox the algorithm defines automatically a $\lambda_{\max}$ and gives the user the possibility of choosing the value of λ between [0,1] which

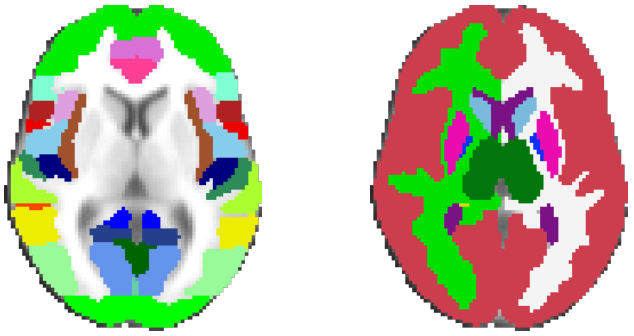
will define the contribution of λ_{max}. Higher values of λ will create more lines with zeros and consequently select less features (more sparse model), whereas lower values of λ will lead to the selection of many features. Consequently, given that the main objective is to reduce the dimensionality problem, ideally higher values of λ are preferred.

Different values of λ were tested to determinate which interval of values would be more adequate. When using intervals with smaller values (between 0.01 and 0.1) the processing data time was much higher as the algorithm selected greater number of features and therefore made the classification procedure very time consuming. For values higher than 0.5 the algorithm selected too few features. Thus, the values which seemed appropriate were in the interval {0.1, 0.2, 0.3, 0.4, 0.5} in order to get a trade off between number of important features selected and computation cost.

Except for the AD vs CN classification, all other groups involved imbalanced data, i.e. one of the classes had much more subjects than the other, consequently the classifier used may favor the larger class. The imbalanced groups in this study had more negative samples (MCI, MCI-NC) than positive samples which could lead therefore to very high values of specificity, as it will have many true negatives (TN) and very few false positives (FP); and very low values of sensitivity, because of the presence of many false negatives (FN) and lower values of true positives (TP) (check equation presented in figure 1.6). The LASSO feature selection algorithm from SLEP was implemented in a way that it can deal well with the imbalanced data as it has a parameter which the user can adjust in order to correct the imbalanced problem, giving higher weights for the class which had less samples. Therefore, for the single modality classification procedure, the imbalance could, in some way, be controlled and avoid really low sensitivity values. However, the group LASSO has not added this ability yet, thus, for the multimodal case the number of samples from the class which was in majority was reduced by selecting less subjects from the ones available, so both classes could have same number of samples.

In order to correctly identify which brain regions were selected in feature selection step, an atlas from the same MNI152 space, where the MRI images were registered, is required. Harvard-Oxford cortical and subcortical structural atlases were the ones used for this step (University of Oxford), covering 48 cortical and 21 subcortical structural areas. This atlas was obtained from T1-weighted images of 21 healthy male and 16 healthy female subjects (18-50 years old). The images were individually segmented by

the Harvard Center for Morphometric Analysis and registered to MNI152 space using FLIRT functionality from FMRIB Software Library (FSL) developed by the Functional MRI Brain (FMRIB) Analysis group from University of Oxford. The two atlas are presented in figure 4.2.



(a) Cortical atlas (b) Subcortical atlas

Figure 4.2: Harvard-Oxford cortical (a) and subcortical (b) structural atlases generated by averaging images to MNI152 space.

4.2.3 Classifiers parameters and evaluation

All the classifiers used in this work were implemented in MATLAB. To perform classifications with the SVM classifier the LIBSVM toolbox was used. Before incorporating the data matrix in this toolbox a feature scaling process is necessary to standardize the range of the features. In this case, the range for the features chosen was [0,1], this was achieved by using the Min-Max scaling method, also known as data normalization. The general formula to use this method is presented in equation 4.1 and was implemented for all the training and test sets from the cross-validation. This dataset scaling is a common requirement before applying SVM and is highly suggested by the LIBSVM guide (Hsu et al., 2003). The advantage of having this bounded range, is that it provides smaller standard deviations, which can suppress the effect of outliers because it will avoid values in greater numeric ranges dominating those in smaller numeric ranges as explained in the guide. Another advantage, which is also referred in the LIBSVM guide, is that it can avoid numerical difficulties during the calculation of the kernels. As kernel values depend on the inner products of feature vectors, the presence of large values might cause numerical problems.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{4.1}$$

By using the linear kernel, it was possible to transform the data matrix into a smaller dimension as explained in section 3.2.1.1 therefore reducing the computational cost of the problem. To determine the optimal C parameter, i.e. the one which is better at accurately predicting unknown data (testing data) a grid search using a exponential growth for the C value is performed, these included 2^k values for $k \in \{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$ in a nested cross-validation procedure, with 10-fold for the outer loop and 5-fold for the inner loop.

To use the LR as the classifier, the procedure is very straightforward, as SLEP toolbox provided all the necessary parameters needed, more precisely, the vector of the features weights $\mathbf{w}$ and the bias term b. These values are then used in equation 3.14.

For both SVM and LR to join information from PET and MRI data a concatenation of the feature vectors was performed. Hence, the feature vectors would have first, features from one modality followed by features from the other modality. For further transformation into a given class label it was used 0.5 as the threshold value.

The GPML toolbox was used to implement the GP classifier. Like in SVM, this classifier also uses kernels, more specifically it can also use linear kernels. These will be the covariance functions which define the classifier. Thus, the algorithm performs the dot product of the feature vectors of the i-th subject and j-th subject, plus a single hyperparameter representing the bias term. For the approximate inference algorithm, the Laplace Approximation was the preferred method.

To join information from different modalities for the GP case, the algorithm from GPML has the possibility of automatically allocating a kernel for each modality and use a scaling hyperparameter representing the modality's weight in the overall kernel, like is represented in equation 4.2 with θ_{MRI} and θ_{PET} representing the weight provided to each modality and b representing the bias in the combined kernel. Thus, taking into consideration equation 3.19, θ is now a set of three hyperparameters ($\theta_{\text{MRI}}, \theta_{\text{PET}}$ and b) which are learnt from the training dataset by type-II maximum likelihood. This

enables automatically setting the kernel weights without needing any grid search with cross-validation, like is usually done in SVM.

$$K_{i,j} = \theta_{\text{MRI}}(X_{\text{MRI},i} \cdot X_{\text{MRI},j}) + \theta_{\text{PET}}(X_{\text{PET},i} \cdot X_{\text{PET},j}) + b \tag{4.2}$$

For the single modality, as it used the imbalanced data, adjustments on the classifiers SVM and GP were also needed. With SVM, the adjustment was done by changing a weight parameter which could give different weights for the different classes. For the GP classifier, the modification was done in the threshold value. Instead of using the usual value 0.5, the threshold was defined by number of the minority class divided by the majority class, this technique of adjusting the threshold to overcome the imbalanced data problem was tested in (Chen et al., 2006) for classification done with Logistic Regression. Furthermore, instead of evaluating the performance of the classifier only using the accuracy, a better statistical value, when having imbalanced data, is evaluating the classification by the balanced accuracy (Koikkalainen et al., 2016) which is the mean of the sensitivity and specificity.

Finally, to evaluate the performance of these different classification methods, 10-fold cross-validation strategy was used. So the data was divided into 10 different training and testing sets taking into attention that these had the same proportion of class positive and class negative from the original data. These 10 different training and testing sets were then provided for all the three classifiers in order to give exactly the same input and to compare the results in a more rigorous way. The process of training and testing is repeated 10 times and then the results are averaged. The averaged accuracy obtained from the testing set will reflect the performance on classifying an unknown and independent dataset.

4.3 Results and Discussion

The analysis of the results will start with the interpretation of each classifier and its relation with the feature selection method used. Then the three classifiers will be compared among them. Furthermore, the posterior probabilities from LR and GP will be

interpreted and compared. Finally, the brain regions which were demonstrated as being relevant for the four different classifications will also be highlighted and compared to the results obtained in previous studies.

4.3.1 Logistic Regression Results

The classification results obtained with LR are presented in figures 4.3, 4.4, 4.5, 4.6 for all 4 groups AD vs CN, AD vs MCI, CN vs MCI and MCI-C vs MCI-NC respectively, for the single and multimodal cases. The blue line representing the accuracy is not visible in some graphs because the balanced accuracy is equal to the accuracy in those cases and therefore is overlying the accuracy line.

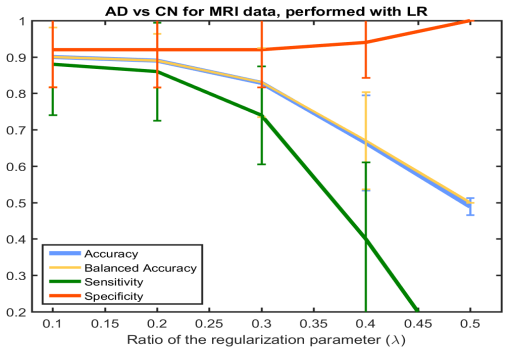
By analysing the results when using MRI data alone one can verify, from the figures, that for higher values of λ the classification results drop drastically, especially in the AD vs CN and AD vs MCI classifications (figure 4.3a and 4.4a) but also happens for the remaining cases (figure 4.5a and 4.6a). On the other side, for PET data this situation does not occur, hence, even if the LR classifier is provided with less features it can obtain good classification results. Nevertheless, with MRI data it is also possible to obtain good results, for example when comparing the results obtained from $\lambda = 0.1$, both MRI and PET graphs show similar and good results. This demonstrates that for the LR classifier when using MRI data, having more features selected is preferred so the classifier can distinguish with higher accuracy the different groups.

For the multimodal test the results were much better than when using just MRI data and similar or a bit better than using just PET data alone, when analysing all values of λ. In this multimodal case, the sensitivity and specificity results also improved, in the sense that the difference between sensitivity and specificity was much lower in comparison to the single modality approach, mainly due to the fact that in the multimodal case the training and testing groups were balanced.

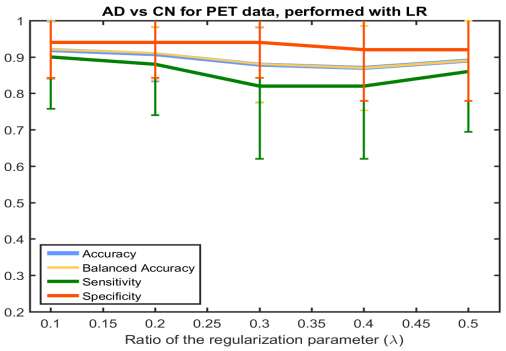
Interestingly, it is possible to see from figure 4.6 that the accuracy and balanced accuracy results for MCI-C vs MCI-NC are not very poor or proximal to chance for both single modal approaches and even show results proximal to AD vs MCI (figure 4.4) and CN vs MCI (figure 4.5). This could be explained by the fact that the selected MCI-C subjects were from a 24 moth period, which is not a very extended conversion period. Therefore,

these MCI-C subjects may have some brain changes similar to the ones found in the AD group, and consequently the classification would be improved.

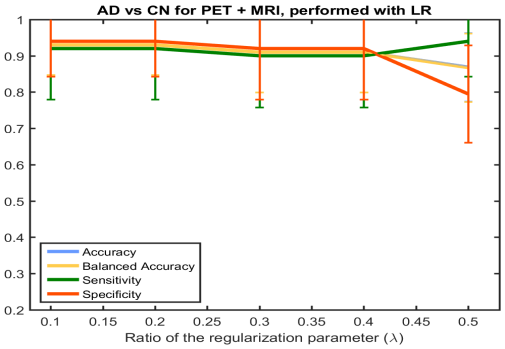
The best results for MRI, PET and multimodal, provided by the optimal λ are presented in table 4.2, and will be further discussed.



(a) MRI data

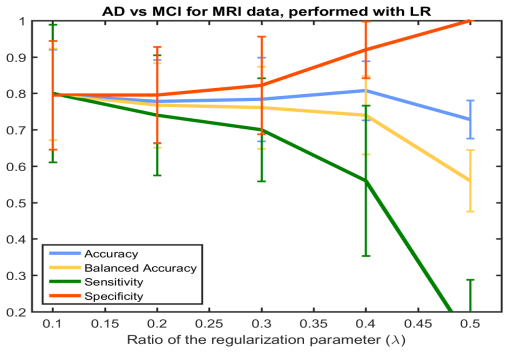


(b) PET data

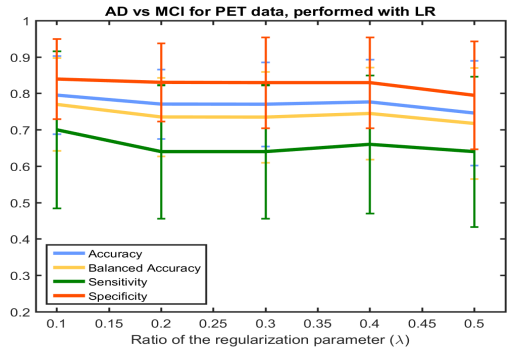


(c) Multimodal (PET + MRI)

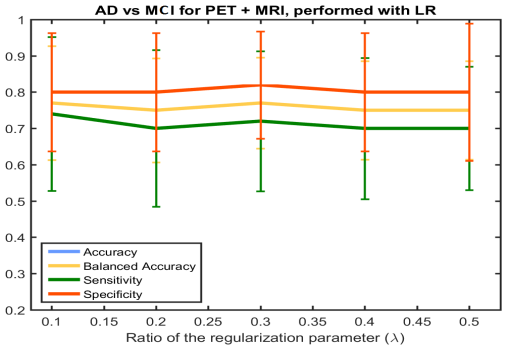
Figure 4.3: AD vs CN classification results using LR as classifier.



(a) MRI data

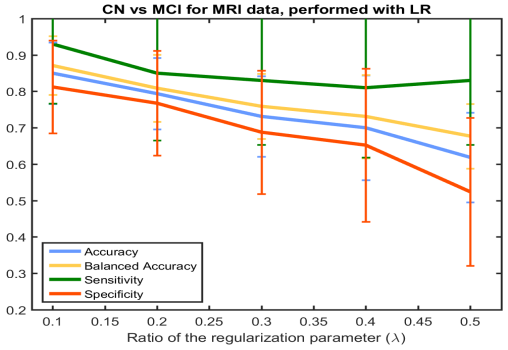


(b) PET data

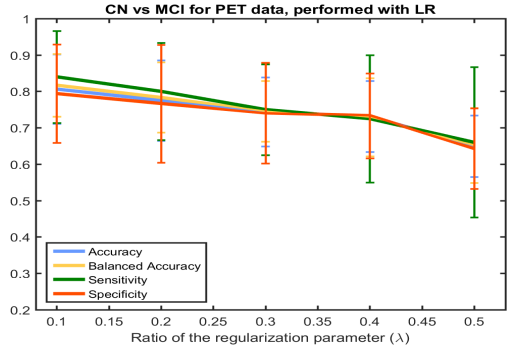


(c) Multimodal (PET + MRI)

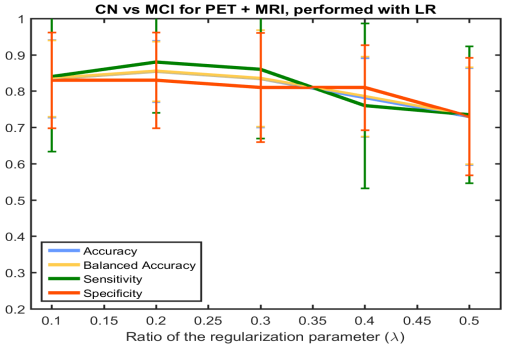
Figure 4.4: AD vs MCI classification results using LR as classifier.



(a) MRI data

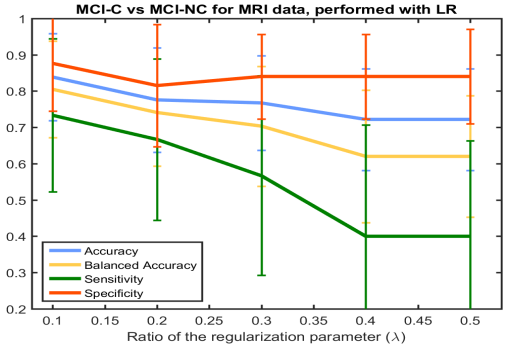


(b) PET data

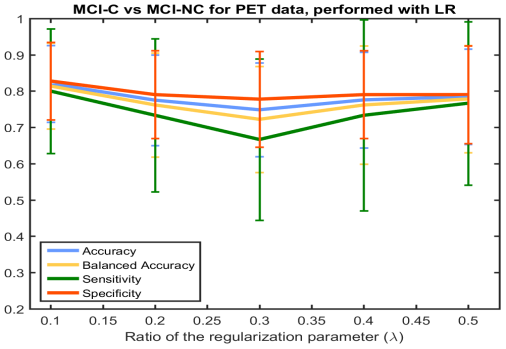


(c) Multimodal (PET + MRI)

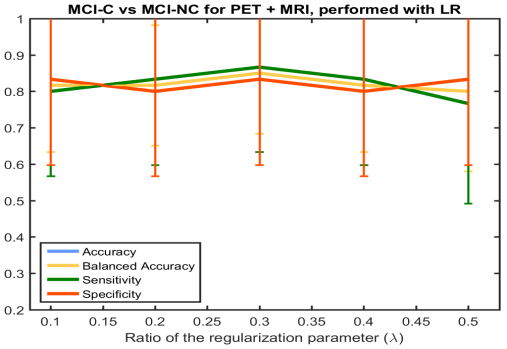
Figure 4.5: CN vs MCI classification results using LR as classifier.



(a) MRI data



(b) PET data



(c) Multimodal (PET + MRI)

Figure 4.6: MCI-C vs MCI-NC classification results using LR as classifier.

4.3.2 Gaussian Process Results

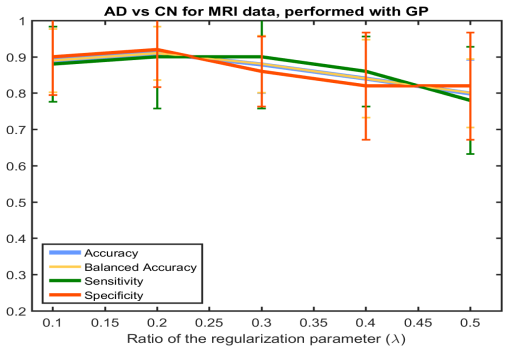
The classification results obtained with GP are presented in figures 4.7, 4.8, 4.9, 4.10 for all 4 groups AD vs CN, AD vs MCI, CN vs MCI and MCI-C vs MCI-NC respectively, for the single and multimodal cases. The blue line representing the accuracy is again not visible in some graphs because of the overlying of the balanced accuracy line.

With GP classifier the trend of descending classification results for higher values of λ, when using just MRI data, is not present except for the MCI-C vs MCI-NC classification (figure 4.10a) that inevitably drops when less features are selected.

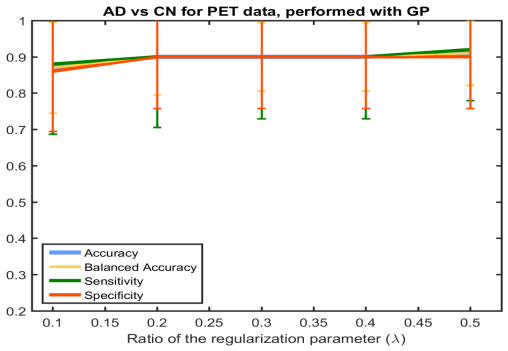
One can also notice, in the single modal results for the imbalanced data (AD vs MCI figures 4.8a and 4.8b; CN vs MCI figures 4.9a and 4.9b,) that, although an adjustment of the threshold value was made, for some λ values the imbalanced problem is not so well controlled and for that reason the graphics show very high values of specificity and low values of sensitivity.

For the multimodal case taking into account the global results (i.e. all λ values), it is possible to verify that for some of the λ values which showed lower performances in the single modal results, the multimodal results could outperform. Nonetheless, comparing with the best results obtained in single modal approach, the multimodal was just better in CN vs MCI and MCI-C vs MCI-NC. This comparison is more noticeable in table 4.2, where the best results for MRI, PET and multimodal, provided by the optimal λ are presented. One can also notice that the results of multimodal do not vary much with the different values of λ.

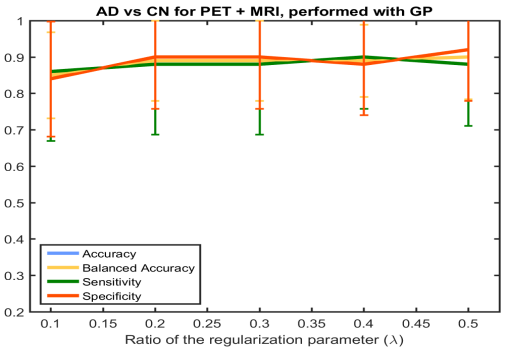
For GP the best classification results are mainly obtained with higher values of λ, which is a plus point as it consequently involves less computation time in classification procedure than when lower values of λ are used.



(a) MRI data

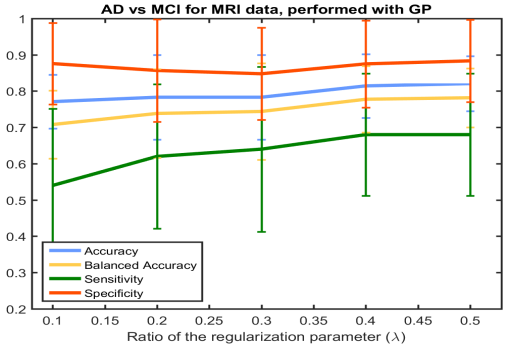


(b) PET data

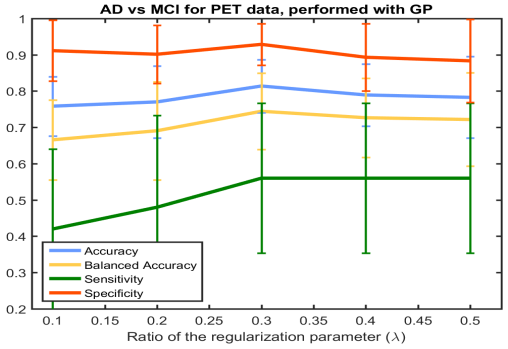


(c) Multimodal (PET + MRI)

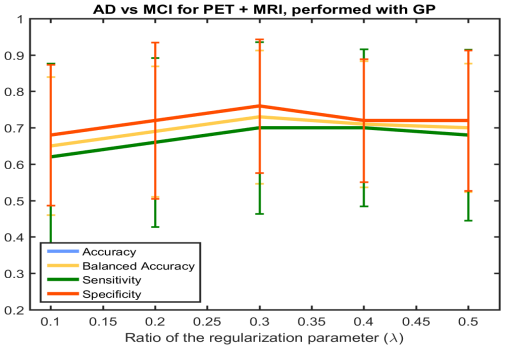
Figure 4.7: AD vs CN classification results using GP as classifier.



(a) MRI data

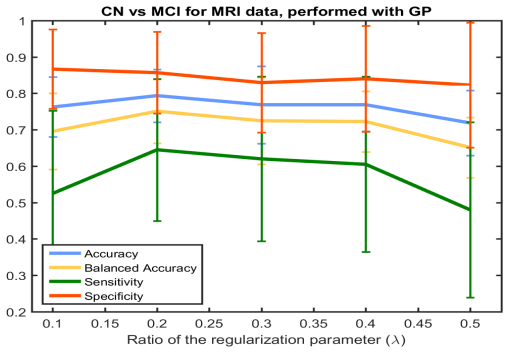


(b) PET data

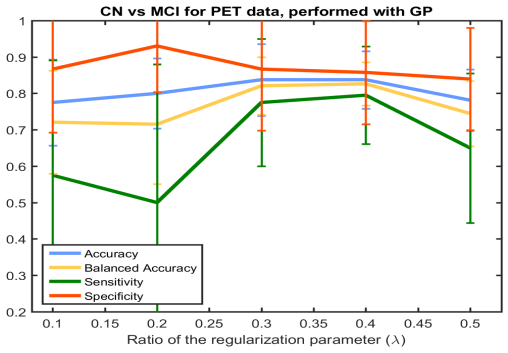


(c) Multimodal (PET + MRI)

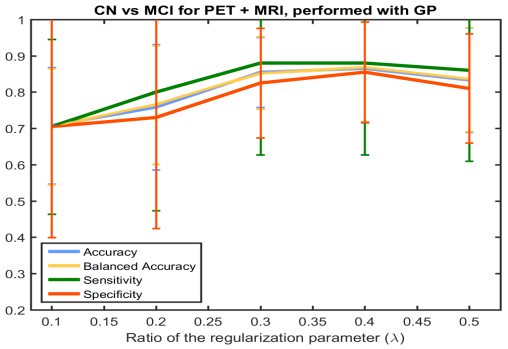
Figure 4.8: AD vs MCI classification results using GP as classifier.



(a) MRI data

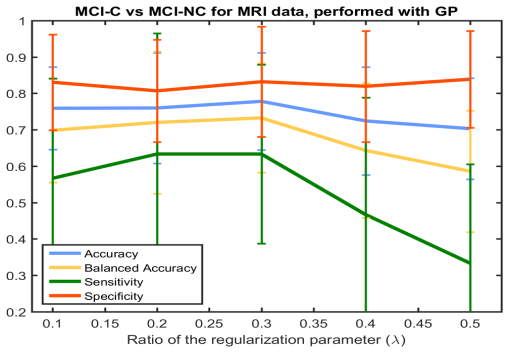


(b) PET data

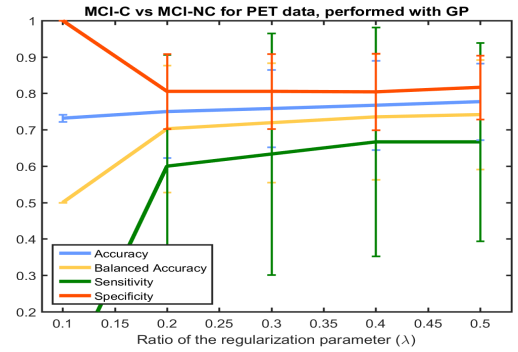


(c) Multimodal (PET + MRI)

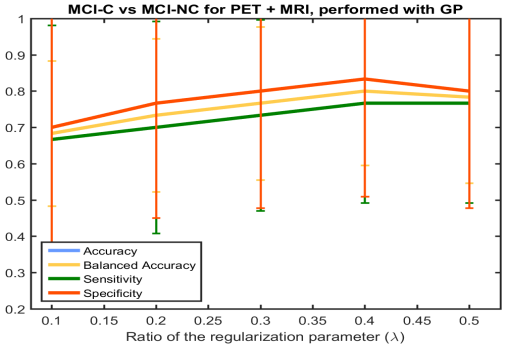
Figure 4.9: CN vs MCI classification results using GP as classifier.



(a) MRI data



(b) PET data



(c) Multimodal (PET + MRI)

Figure 4.10: MCI-C vs MCI-NC classification results using GP as classifier.

4.3.3 SVM Results

The classification results obtained with SVM are presented in figures 4.11, 4.12, 4.13, 4.14 for all 4 groups AD vs CN, AD vs MCI, CN vs MCI and MCI-C vs MCI-NC respectively, for the single and multimodal cases. Likewise in previous results some of these graphs the blue line from the accuracy is not visible due to the balanced accuracy line.

For both single modal classifications done with SVM for AD vs CN (figure 4.11) the best results are achieved for lower values of λ but higher values can also present good performance. The results obtained with both modalities together (figure 4.11c) independently of λ value do not vary much and do not outperform the single modal results.

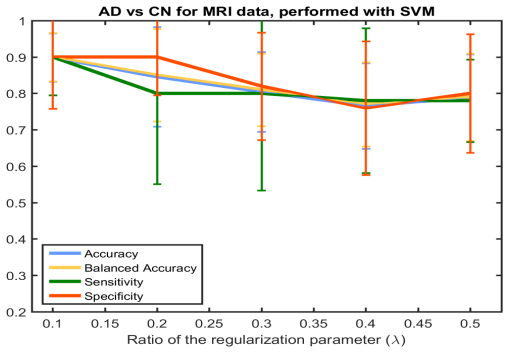
It seems that for higher values of λ the imbalanced issue for the single PET and MRI classifications is well controlled in most cases (figures 4.12a, 4.13a, 4.13b), except for AD vs MCI with PET data (figure 4.12b), that presents high specificity values but low sensitivity values, which means that there are many false negative samples.

Likewise LR and GP, with SVM for MCI-C vs MCI-NC classification obtained with MRI data (figure 4.14a) the results also drop for higher values of λ. While with PET data (figure 4.14b) this does not occur and even improve the imbalance problem as sensitivity results start rising.

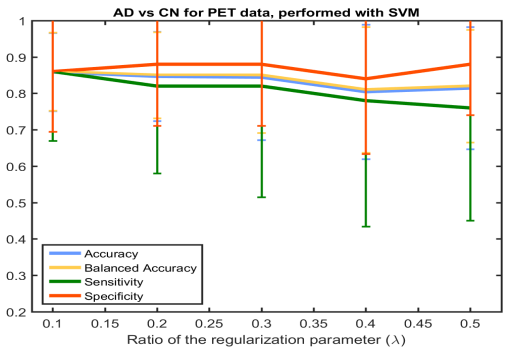
Expect for AD vs CN, the multimodal approach did show more variation in the results obtained for different λ values. For AD vs MCI (figure 4.12c) the best results are obtained with $\lambda = 0.5$ but for CN vs MCI and MCI-C vs MCI-NC (figures 4.13c and 4.14c) intermediate values of λ are preferred.

Similarly as GP, SVM can also retrieve good classifications for higher λ values and therefore reduce the computation time in the training procedure.

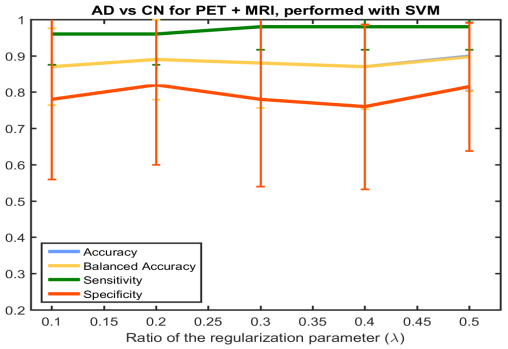
The best results obtained, provided by the optimal λ, and in this case, the optimal C parameter, are presented in table 4.2. These will be discussed in the following section and compared with the LR and GP results.



(a) MRI data

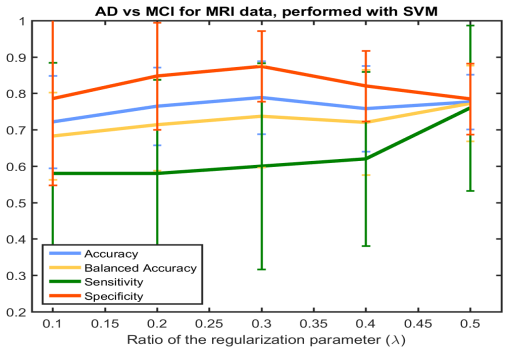


(b) PET data

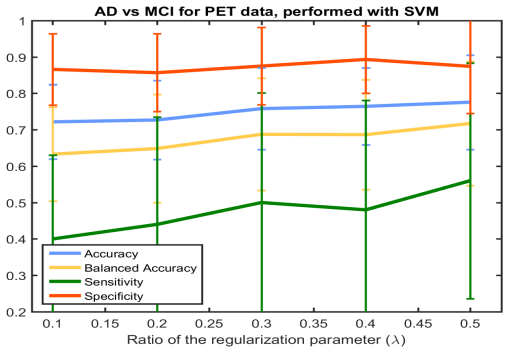


(c) Multimodal (PET + MRI)

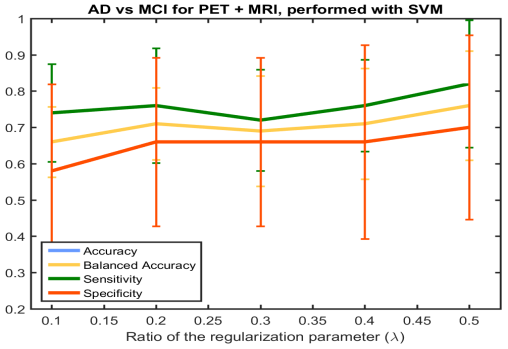
Figure 4.11: AD vs CN classification results using SVM as classifier.



(a) MRI data

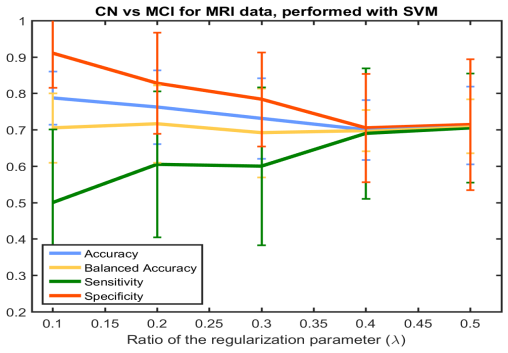


(b) PET data

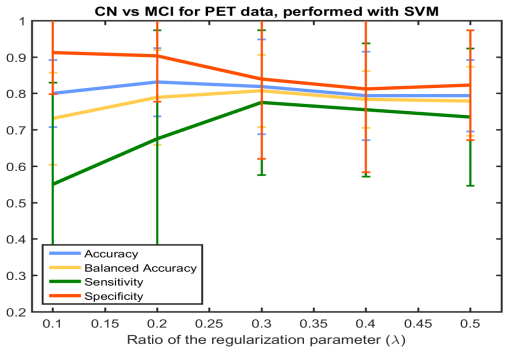


(c) Multimodal (PET + MRI)

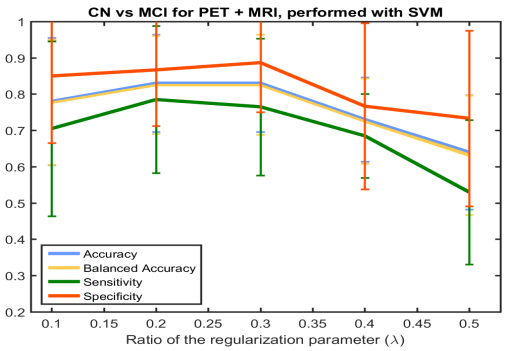
Figure 4.12: AD vs MCI classification results using SVM as classifier.



(a) MRI data

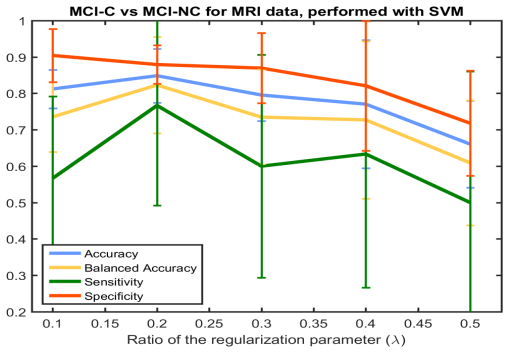


(b) PET data

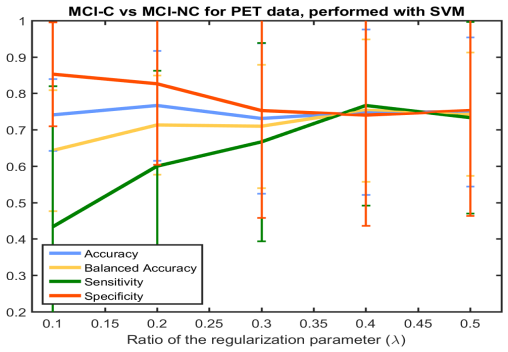


(c) Multimodal (PET + MRI)

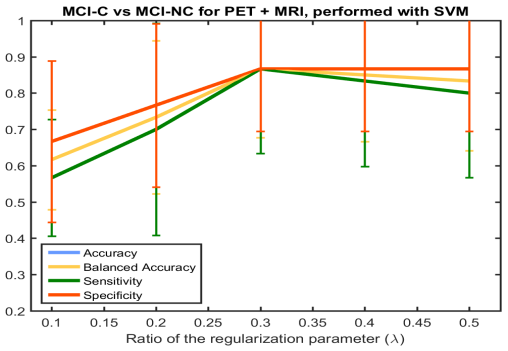
Figure 4.13: CN vs MCI classification results using SVM as classifier.



(a) MRI data



(b) PET data



(c) Multimodal (PET + MRI)

Figure 4.14: MCI-C vs MCI-NC classification results using SVM as classifier.

4.3.4 LR, GP and SVM results comparison

Table 4.2 presents the best results obtained for each classifier with the optimal λ value (and optimal C parameter for SVM) for a better performance comparison between the three classifiers. One can see from this table that lower values of λ, more precisely 0.1, 0.2 and 0.3, are preferred when using LR as classifier. On the other hand, GP and SVM classifiers can achieve better results with higher values of λ. Which means that they do not need so many features selected as LR does.

For AD vs CN the results obtained with the three classifiers are very similar and are near 90% accuracy. Using the multimodal approach, in this case, did not improve the single modality results for SVM or GP, just for the LR classifier.

Looking at the AD vs MCI results, if only evaluating the accuracy, one could conclude that GP is preferred for better classification results but as the data is imbalanced for both single modal cases it is more adequate to evaluate the classification with the balanced accuracy. Thus, it appears that the best classifier is actually LR. The multimodal classification for AD vs MCI also did not outperform the single modal classification results.

For CN vs MCI classification, when using just MRI data, the LR could get better results then SVM and GP. However, for PET and multimodal results, GP outperformed SVM and slightly LR results. In this case the multimodal approach could show improvement in results.

Interestingly, although LR and GP mostly showed better results than SVM, for the MCI-C vs MCI-NC classification, SVM got much better results than GP and slightly higher accuracies than LR. Except for PET single modal test, where LR was the best at differentiating these two groups. In this case the multimodal approach, regardless of the classifier used, did improve the single modal results. Showing therefore that is advantageous to use a multimodal data to distinguish MCI-C from MCI-NC.

As stated previously, the C parameter was tested for 2^k values with $k \in \{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$. The results obtained for all these values, showed that SVM had similar classification performances. If the higher C parameter did not improve the results in comparison to the lower values of C, it is preferred to choose the lower values for C parameter because these will provide a larger margin and therefore create a model with

more ability to generalize the classifier to unseen data. Besides, smaller C values have lower computation time. Therefore, in table 4.2 the C parameter chosen is the lower value which presented better classification performance.

			SVM	GP	LR
AD vs CN	Accuracy ± std (%)	MRI	89.9 ± 6.7	**91.0 ± 7.4**	90.0 ± 8.2
		PET	85.9 ± 10.7	91.0 ± 8.8	**91.9 ± 7.9**
		Multimodal	89.9 ± 9.4	90.0 ± 11.6	**93.0 ± 8.3**
	Balanced Accuracy ± std (%)	MRI	90.0 ± 6.7	**91.0 ± 7.4**	90.0 ± 8.2
		PET	86.0 ± 10.8	91.0 ± 8.8	**92.0 ± 7.9**
		Multimodal	89.8 ± 9.5	90.0 ± 11.6	**93.0 ± 8.3**
	λ value (# features)	MRI	0.1 (139937)	0.2 (38312)	0.1 (319777)
		PET	0.1 (319777)	0.1 (319777)	0.1 (319777)
		Multimodal	0.5 (1788)	0.5 (1788)	0.2 (107620)
	C-value	MRI	0.0313	-	-
		PET	0.0625	-	-
		Multimodal	0.0313	-	-
AD vs MCI	Accuracy ± std (%)	MRI	77.7 ± 7.5	**82.1 ± 7.5**	79.6 ± 12.4
		PET	77.6 ± 13.0	**81.4 ± 7.3**	79.5 ± 10.7
		Multimodal	76.0 ± 15.1	73.0 ± 18.3	**77.0 ± 12.5**
	Balanced Accuracy ± std (%)	MRI	77.2 ± 10.4	78.2 ± 8.1	**79.8 ± 12.6**
		PET	71.7 ± 17.0	74.4 ± 10.5	**77.0 ± 12.7**
		Multimodal	76.0 ± 15.1	73.0 ± 18.3	**77.0 ± 12.5**
	λ value (# features)	MRI	0.5 (1384)	0.3 (103252)	0.1 (358995)
		PET	0.5 (19597)	0.3 (103252)	0.1 (358995)
		Multimodal	0.5 (13096)	0.3 (81360)	0.3 (81360)
	C-value	MRI	2	-	-
		PET	0.0313	-	-
		Multimodal	0.0313	-	-
CN vs MCI	Accuracy ± std (%)	MRI	76.3 ± 10.1	79.4 ± 7.3	**85.0 ± 8.4**
		PET	81.9 ± 13.0	**83.8 ± 7.9**	80.6 ± 9.5
		Multimodal	83.0 ± 13.4	**86.6 ± 15.0**	85.4 ± 8.5
	Balanced Accuracy ± std (%)	MRI	71.7 ± 10.7	75.1 ± 8.8	**87.1 ± 8.1**
		PET	80.7 ± 10.0	**82.6 ± 6.0**	81.7 ± 8.6
		Multimodal	82.6 ± 13.5	**86.8 ± 14.9**	85.5 ± 8.2
	λ value (# features)	MRI	0.2 (48813)	0.2 (48813)	0.1 (173172)
		PET	0.3 (52737)	0.4 (12674)	0.1 (303954)
		Multimodal	0.3 (49442)	0.4 (9888)	0.2 (155631)
	C-value	MRI	0.0313	-	-
		PET	0.0313	-	-
		Multimodal	0.0313	-	-
MCI- C vs MCI-NC	Accuracy ± std (%)	MRI	**84.9 ± 7.4**	77.8 ± 13.3	83.9 ± 12.0
		PET	74.9 ± 22.7	77.7 ± 10.5	**82.1 ± 10.6**
		Multimodal	**86.7 ± 18.9**	80.0 ± 20.5	85.0 ± 16.6
	Balanced Accuracy ± std (%)	MRI	**82.3 ± 13.2**	73.3 ± 15.0	80.5 ± 13.3
		PET	75.4 ± 19.6	74.2 ± 15.0	**81.4 ± 11.8**
		Multimodal	**86.7 ± 18.9**	80.0 ± 20.5	85.0 ± 16.6
	λ value (# features)	MRI	0.2 (40777)	0.3 (40777)	0.1 (1594449)
		PET	0.4 (27642)	0.5 (8227)	0.1 (317776)
		Multimodal	0.3 (38018)	0.4 (11928)	0.3 (38018)
	C-value	MRI	0.0313	-	-
		PET	0.0313	-	-
		Multimodal	0.0313	-	-

Table 4.2: Best classification results obtained with SVM, GP and LR.

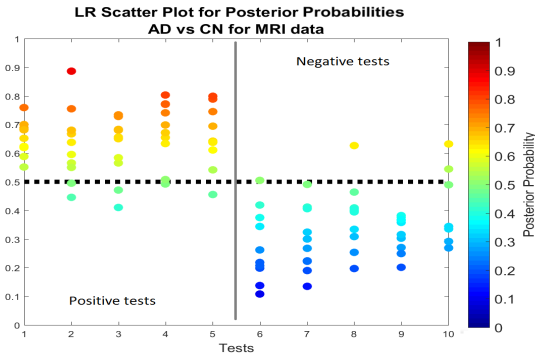
4.3.5 Posterior Probabilities

As GP and LR are probabilistic classifiers their predicted posterior probabilities should be evaluated. Thus, a scatter plot of the posterior probabilities obtained with these two classifiers are presented in 4.15, 4.16, 4.17 and 4.18 for the LR and in 4.19, 4.20, 4.21 and 4.22 for GP. These figures plot the posterior probabilities obtained for the test samples during cross-validation. When a given test sample belongs to the positive class the posterior probability of that sample should be near 1, if the test sample belongs to the negative class then it should be near 0. The colors in these images help to distinguish if the posterior probability obtained is near 1 or 0, by matching to the colors presented in the colorbar scale on the right side of each image. The colors in the middle (light green/light blue) are the ones that indicate that the posterior probabilities are near the threshold value, which is represented by the horizontal dashed line, meaning that for those samples the classifier has more uncertainty about the prediction made. Colors near dark red or near dark blue indicate that the classifier is almost sure that the sample belongs to the positive class or to the negative class, respectively.

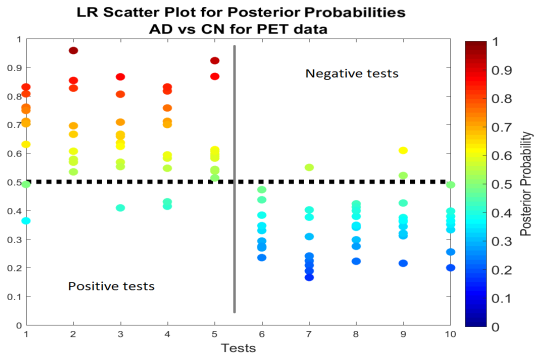
By comparing the results obtained with LR and GP it is very clear that GP can present posterior probabilities with more confidence than LR as it presents more points near the extreme values (1 or 0). From a clinical point of view, analysing just the posterior probabilities obtained with LR, as it has many samples which lie near the threshold value, probably it would be necessary to perform another test or add another modality to see if this posterior probability could be turned into a value which is near 1 or 0. On the other hand, using the posterior probabilities obtained with GP may be more relevant in a clinical point of view, in order to help choose treatment procedure with more confidence in the treatment chosen being adequate for that patient. Therefore, analysing the classification results for example in the MCI-C vs MCI-NC classification, although LR could get better accuracy and balanced accuracy results than GP, as shown in table 4.2, comparing the posterior probabilities from LR in figure 4.18 and from GP in figure 4.22, shows that GP can be more certain about the predictions made. This difference found in the posterior probabilities clarifies the importance of identifying these quantities when performing classifications, specially for a clinical purpose to help define the most adequate treatment.

Although GP presented higher levels of certainty for the correct test samples it did not prove to be much better than LR as it also clearly presented a lot more misclassified test samples. If these misclassified samples were near the threshold value it would less concerning, however this did not happen in some cases, and GP showed high confidence even for the misclassified tests. This problem is specially revealed in the AD vs MCI and CN vs MCI for the single modal approach showing many FN and few TP, thus, one could interpret this result as being a consequence of having imbalanced data. This is also in accordance with graphics presented previously (AD vs MCI figures 4.8a and 4.8b; CN vs MCI figures 4.9a and 4.9b,) which showed very high values of specificity and low values of sensitivity.

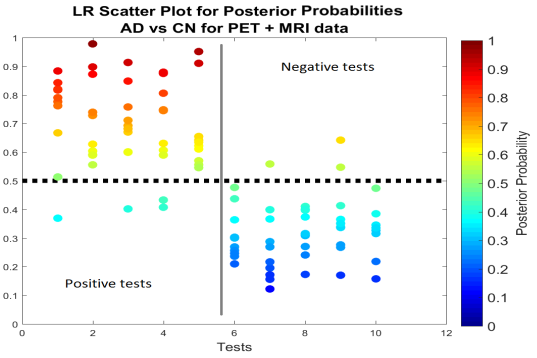
The results of this probabilistic analysis show that having the posterior probabilities could indeed be useful in order to have more information than just a predicted class and presents another way to compare the performance of different classifiers. For example, if one compares two classifiers having same accuracy, sensitivity and specificity, they would be interpreted as having equal performance, nonetheless, looking into the posterior probabilities it is possible to figure out which would be preferred based on the posterior probabilities returned. Thus, it seems quite essential to explore these posterior probabilities provided by the classifiers more deeply and try to identify how correct they are, or how they could be improved.



(a) MRI data

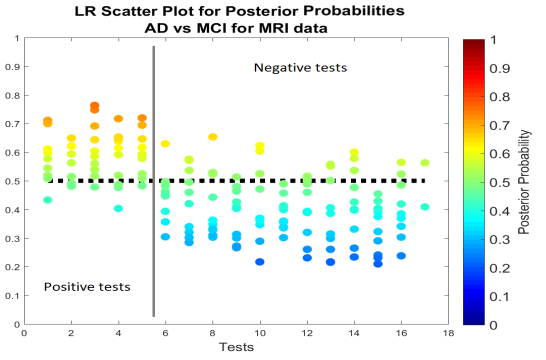


(b) PET data

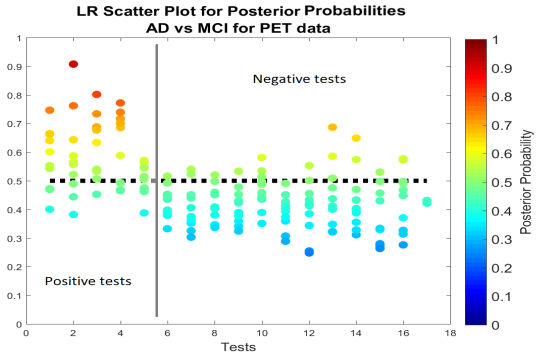


(c) Multimodal (PET + MRI)

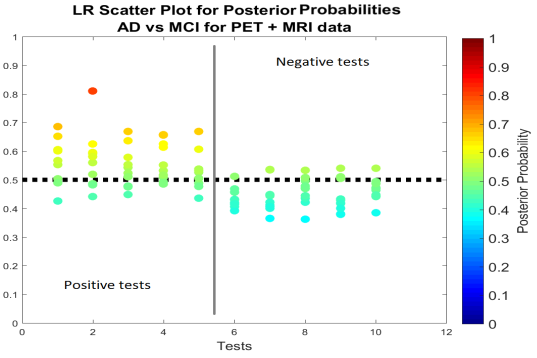
Figure 4.15: Posterior Probabilities obtained with LR for ADvsCN.



(a) MRI data

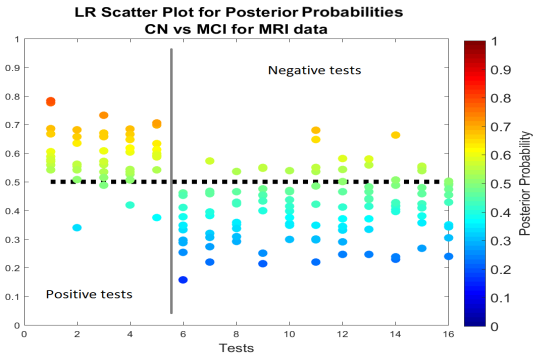


(b) PET data

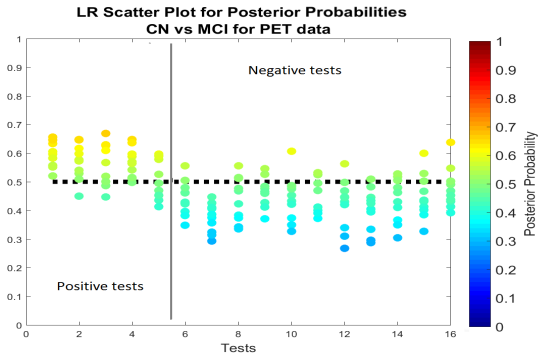


(c) Multimodal (PET + MRI)

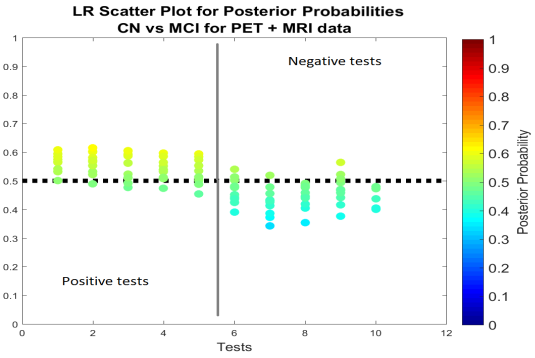
Figure 4.16: Posterior Probabilities obtained with LR for ADvsMCI.



(a) MRI data

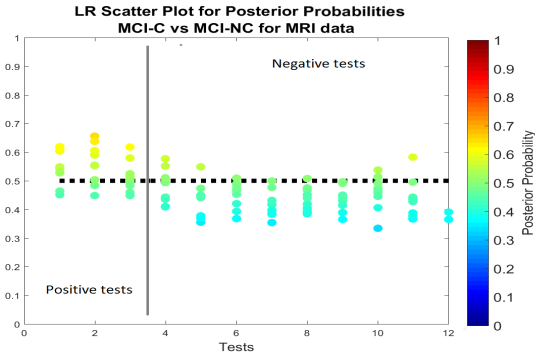


(b) PET data

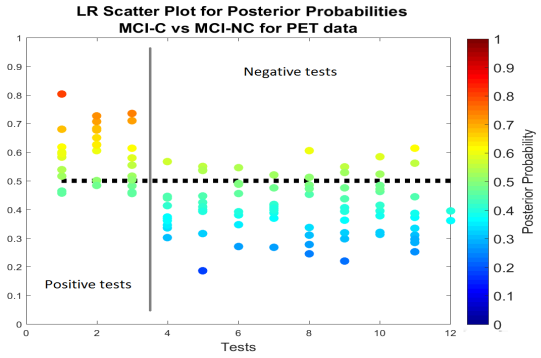


(c) Multimodal (PET + MRI)

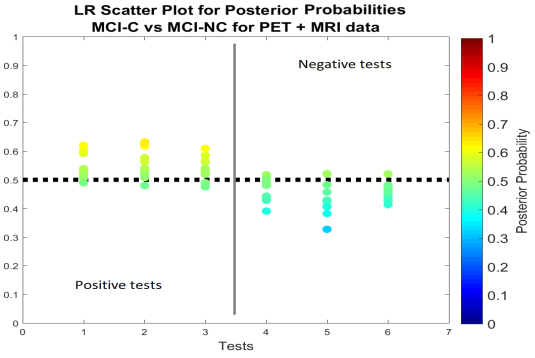
Figure 4.17: Posterior Probabilities obtained with LR for CNvsMCI.



(a) MRI data

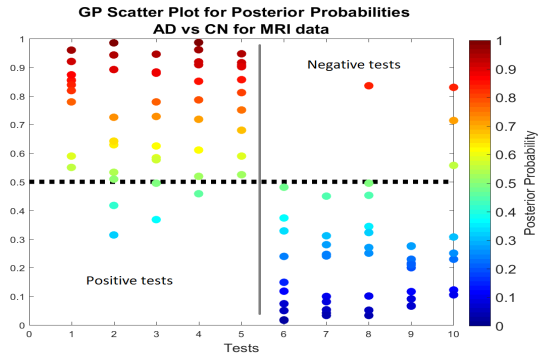


(b) PET data

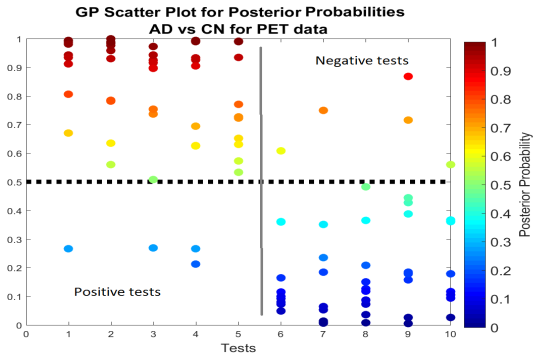


(c) Multimodal (PET + MRI)

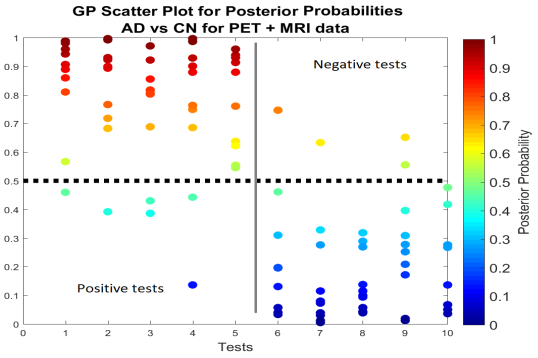
Figure 4.18: Posterior Probabilities obtained with LR for MCI-CvsMCI-NC.



(a) MRI data

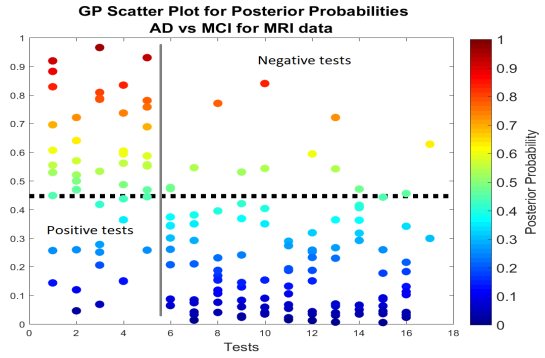


(b) PET data

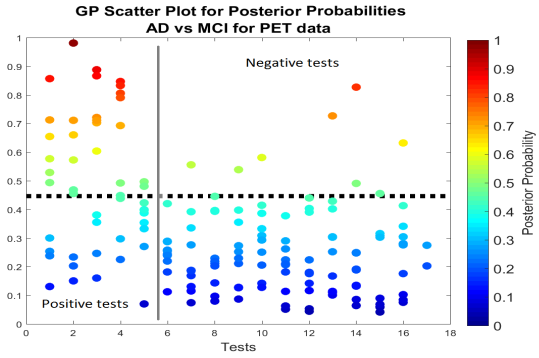


(c) Multimodal (PET + MRI)

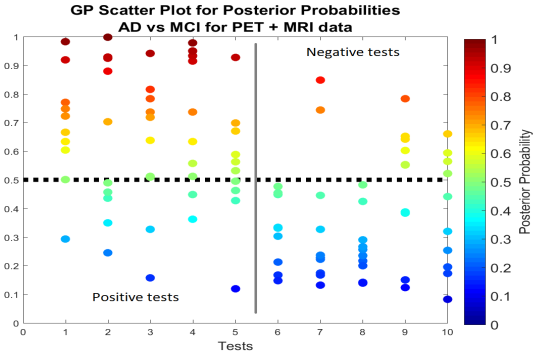
Figure 4.19: Posterior Probabilities obtained with GP for ADvsCN.



(a) MRI data

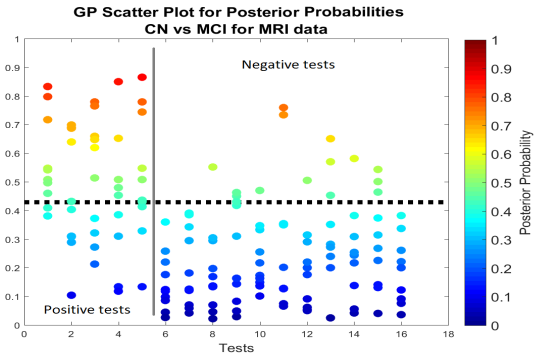


(b) PET data

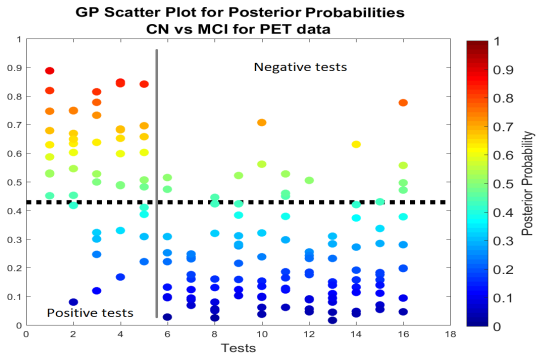


(c) Multimodal (PET + MRI)

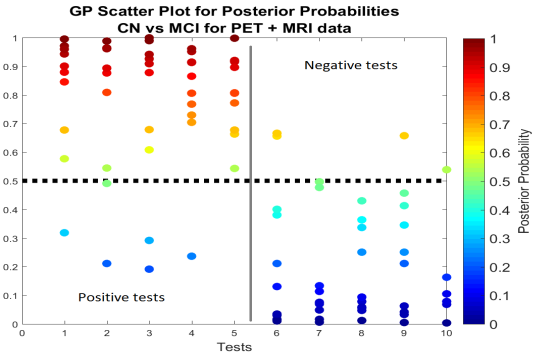
Figure 4.20: Posterior Probabilities obtained with GP for ADvsMCI.



(a) MRI data

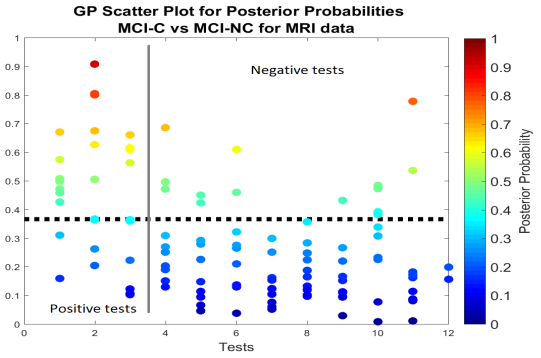


(b) PET data

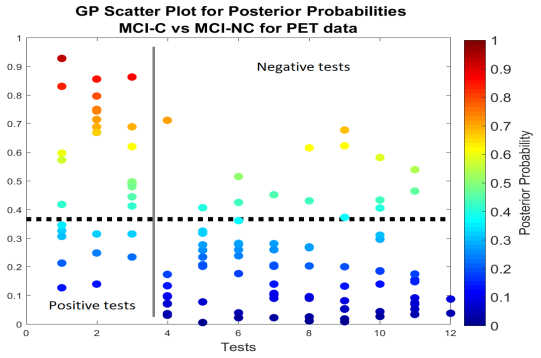


(c) Multimodal (PET + MRI)

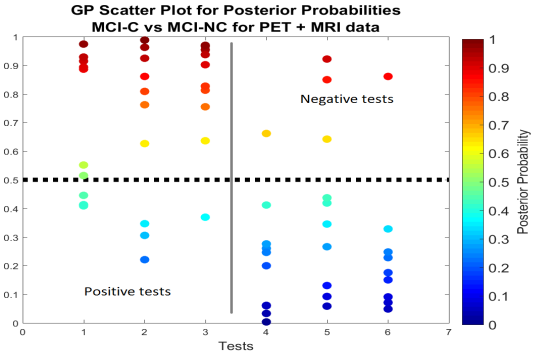
Figure 4.21: Posterior Probabilities obtained with GP for CNvsMCI.



(a) MRI data



(b) PET data



(c) Multimodal (PET + MRI)

Figure 4.22: Posterior Probabilities obtained with GP for MCI-CvsMCI-NC.

4.3.6 Selected Features

By using the features selection algorithm from SLEP it was possible to obtain the most relevant features to distinguish the different classes in this study. The Harvard subcortical atlas was then used to discover which subcortical brain regions were selected as pertinent for better classification results and therefore are seen as relevant in AD studies.

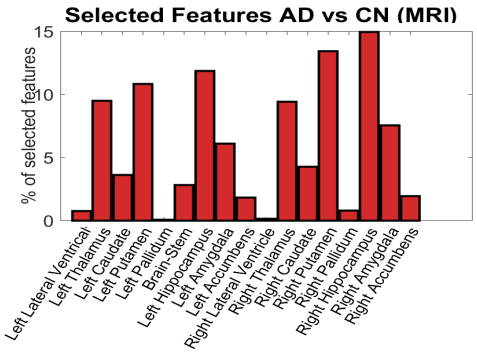
For all the tests, most of the selected features were from cerebral white matter and cerebral grey matter, which is very much expected as these are the areas where most differences are found when comparing to images of PET or MRI, showing differences in metabolic rate and volume shrinkage, respectively. Nonetheless, the algorithm could also distinguish other more specific relevant regions in the subcortical area, some of them can be seen in figures 4.23, 4.24, 4.25 and 4.26 for AD vs CN, AD vs MCI, CN vs MCI and MCI-C vs MCI-NC respectively.

As explained in section 3.1 group LASSO will analyse information from both modalities and try to select the most important features which are in common. One can verify by the graphics that these selected features were not necessarily commonly selected by the two different modalities when it was done independently. This explains why in the group LASSO method very few features were selected. With this result, in one way, group LASSO has the advantage of having the ability of reducing drastically a really high-dimensional problem. On the other side, it has the disadvantage of losing information as many features which could help distinguishing the different classes are not being selected. Therefore, the multimodal results are not so appealing and in many cases single modality results outperformed the multimodal case. The fact that PET images had to be transformed into MRI space could also be a factor affecting this multimodal performance, as they were not originally from the same exact space.

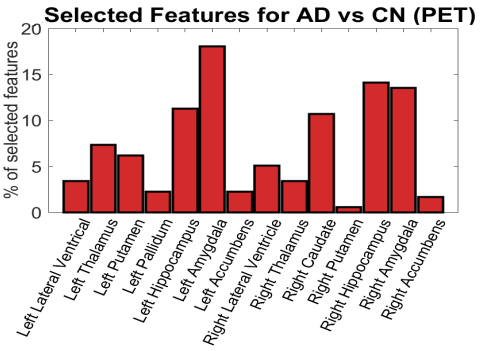
From the selected features in the AD vs CN classification, presented in figure 4.23, one can see that the regions that stand out are the hippocampus and amygdala, which is in accordance with many studies in this area. Both of these regions are located medially within the temporal lobes. As mentioned in chapter 1 the hippocampus is crucial for the formation of short-term memories, and from what is known, the amygdala also has a relation with primary role in memory processing and emotion reactions.

For the AD vs MCI classification, by using just MRI data, the region that was mostly selected among the rest of the subcortical regions was the putamen. This region was found to be related to AD showing strongly reduced volumes in affected patients (De Jong et al., 2008). For PET, besides the the hippocampus and amygdala, the right lateral ventricle is shown to be also relevant to distinguish AD vs MCI patients. This could be explained by the fact that these lateral ventricles have ependyma, a specialised form of epithelium which is involved in the production of CSF but also is shown to serve as a reservoir for neuroregeneration. As stated previously in chapter 1, in an early stage the brain tries to compensate the caused damages. Thus, probably this brain region in MCI patients will have a superior metabolic activity in comparison to AD. Studies evaluating MRI neuroimages have shown that AD patients normally present an enlargement of the lateral ventricles (Frisoni et al., 2010).

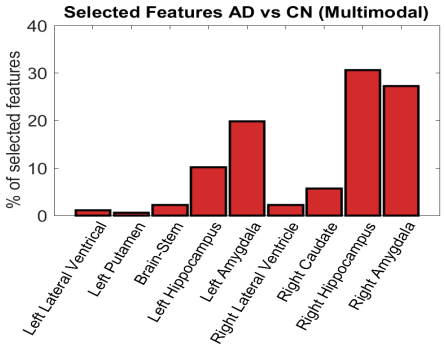
Both CN vs MCI and MCI-C vs MCI-NC classifications highlight, in addiction, the thalamus as an important region of the brain to analyse, when attempting to distinguish between these classes. The thalamus is functionally connected to the hippocampus as part of the extended limbic system with respect to spatial memory, being crucial for human episodic memory, and was already shown as a relevant region in previous studies for early AD (De Jong et al., 2008; Aggleton et al., 2016).



(a) MRI data

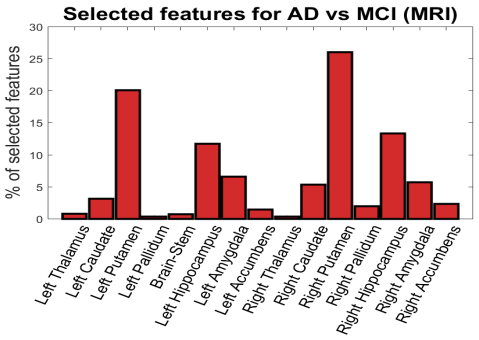


(b) PET data

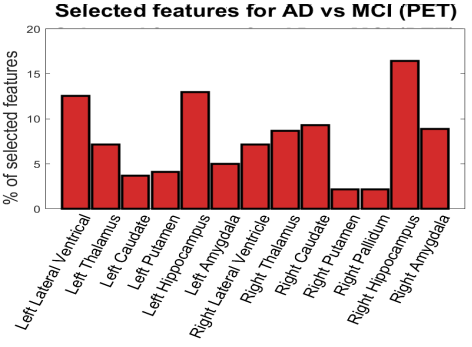


(c) Multimodal (PET + MRI)

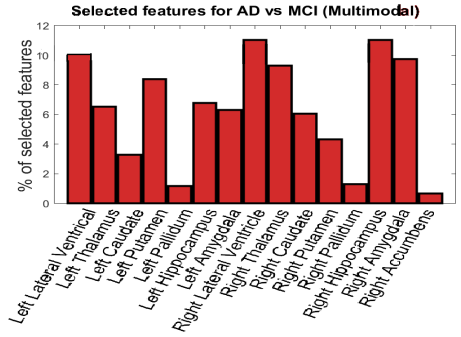
Figure 4.23: Subcortical brain regions selected for ADvsCN.



(a) MRI data

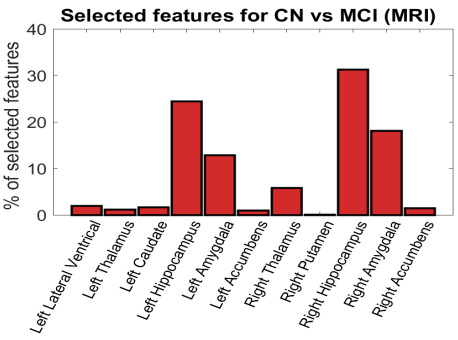


(b) PET data

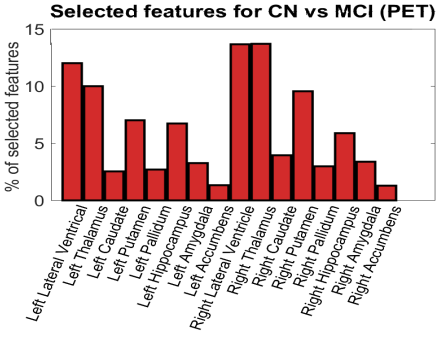


(c) Multimodal (PET + MRI)

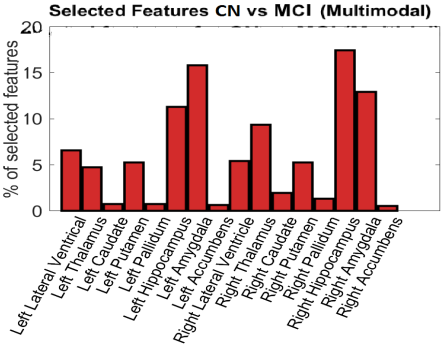
Figure 4.24: Subcortical brain regions selected for ADvsMCI.



(a) MRI data

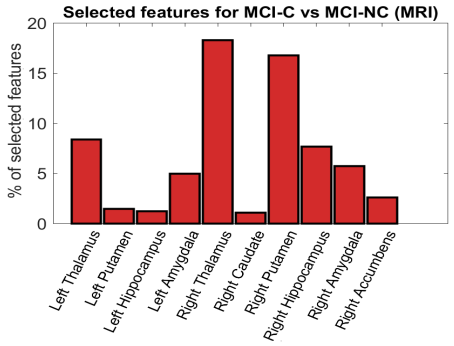


(b) PET data

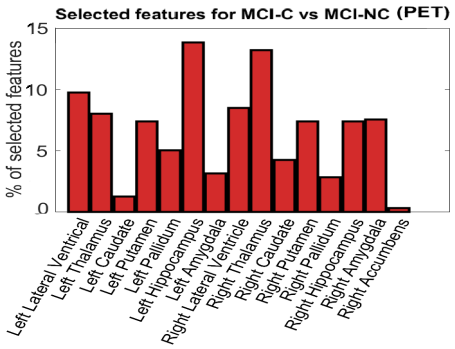


(c) Multimodal (PET + MRI)

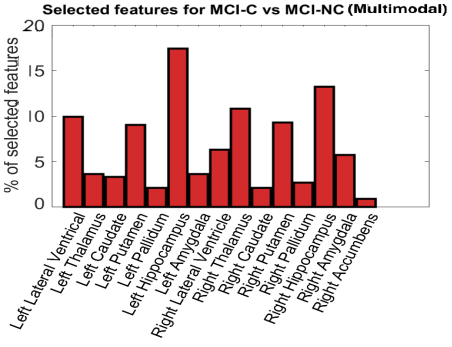
Figure 4.25: Subcortical brain regions selected for CNvsMCI.



(a) MRI data



(b) PET data



(c) Multimodal (PET + MRI)

Figure 4.26: Subcortical brain regions selected for MCI-CvsMCI-NC.

Chapter 5

Conclusions and Future work

In this work three different classifiers (LR, GP and SVM) were tested with two different modalities (MRI and PET) first individually used and then combined, in order to evaluate their performance in distinguishing the different groups (AD, CN, MCI, MCI-C, MCI-NC) from this study, and essentially how they could contribute to detect early cases of Alzheimer by distinguishing MCI-C from MCI-NC .

Group LASSO multi-task selected features based on PET and MRI information, more precisely selected features which were relevant for both modalities. This method proved to be a good method for multimodal use in the sense that it is very useful at reducing the cost computation as it induces sparsity. Moreover, it can be helpful to be sure about specific brain regions which are more prone to brain changes caused by AD. Usually, the motivation for using the Group LASSO multi-task is to select features based on the assumption that all tasks/modalities used share a common sparsity pattern. Nevertheless, as it was possible to see by the single modal approach, MRI and PET do not necessarily select the same features. Consequently, using the Group LASSO multi-task, some important features that are not equally relevant for both modalities may not be selected, which is a limitation of this method. Therefore, a more precise and interesting study would involve adding features which are not common but are at least relevant to one of the modalities. This could improve the classification performance.

In terms of the three classifiers performance, the LR classifier proved to be better at achieving good classification results when more features were selected, while SVM and GP could get reasonable results with less features selected. Nonetheless, although LR

needed more features the computation time was smaller in comparison to GP or SVM. The classifier that was indeed the most computationally expensive was SVM because of the grid search done for the C parameter. Some studies, in order to avoid this issue usually use the default value for this parameter. In most of the classifications tested in this work (AD vs CN, AD vs MCI, CN vs MCI) GP and LR outperformed SVM classification results, and thus show that these classifiers are useful for these type of classification problems and maybe even better than the so widely used SVM. On the other hand, although SVM did not outperform GP or LR in most cases, the results for MCI-C vs MCI-NC classification were better than GP and LR, at least when using MRI and multimodal data. When using PET data, for MCI-C vs MCI-NC classification, LR presented better results.

For the multimodal case better results were expected than when using just one modality. However, this was not the case in some of the experiments, probably because the results obtained with the single modal approach were already very good at least for AD vs CN. Fortunately, for the CN vs MCI and for MCI-C vs MCI-NC classification, the multimodal approach did help improve the classification results and show that for a early diagnosis it is indeed relevant to combine information of different modalities.

By analysing the posterior probabilities of LR and GP one could see clearly that the results obtained with GP were preferred to the ones obtained with LR because it seemed to have more confident results, and therefore it presents a plus point in comparison to LR.

The fact that the number of subjects used in this work was not very high, and in some groups the samples were imbalanced, could be a limitation of this work. On the other hand, having more subjects for all groups would increase the computation time of the training step. A limitation for the multimodal classification results could be the fact that PET and MRI were not originally from the same space and PET data had to be transformed into MRI space.

This work demonstrates the advantage of using the probabilistic classifiers GP and LR as they can achieve state-of-the art classification results and be better than SVM, in some cases, and also because they can provide posterior probabilities that will help evaluate how confident the classifier is on its predictions. It also emphasises the need of a more

complete understanding of how correct these posterior probabilities are in order to make these results even more appealing for clinical use.

In the future, a similar study could also be done by converting the results of SVM into a probabilistic framework by calculating the distance that a determinate sample is to the hyperplane defined by the SVM classifier and translating the output into probability intervals. In addition, using also other modalities like genetic information or CSF, for example, would also be interesting to investigate. Furthermore, besides the 24 months conversion period tested in this work for the MCI subjects, other extended conversion periods could be experimented in a future work.

Bibliography

J.P. Aggleton, A. Pralus, A.J. Nelson, M. Hornberger, et al. Thalamic pathology and memory loss in early Alzheimer's disease: moving the focus from the medial temporal lobe to Papez circuit. *Brain*, 139:1877–1890, 2016. doi: 10.1093/brain/aww083.

MS. Albert, ST. DeKosky, D. Dickson, B. Dubois, HH. Feldman, N. Fox, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7(3):270–279, 2011. doi: 10.1016/j.jalz.2011.03.008.

J. Alom, I. Llinares, and S. Fajardo. Clinical Approach to Diagnosis of Pre-Dementia Alzheimer's Disease. *Dementia and Geriatric Cognitive Disorders EXTRA*, 2(1):332–342, 2012. doi: 10.1159/000341776.

Alzheimer Europe. 2013: The prevalence of dementia in Europe, Portugal. [online], 2013. URL http://www.alzheimer-europe.org/Policy-in-Practice2/Country-comparisons/2013-The-prevalence-of-dementia-in-Europe/Portugal. [Accessed November, 2015].

Alzheimer's Association. Alzheimer's disease facts and figures. *Alzheimer's Dementia: The Journal of the Alzheimer's Association*, 12(4):459–509, 2016. doi: 10.1016/j.jalz.2016.03.001.

C. Ballard, S. Gauthier, C. Brayne, D Aarsland, and E. Jones. Alzheimer's disease. *The Lancet*, 377(9770):1019–1031, 2011. doi: 10.1016/S0140-6736(10)61349-9.

J. Bischkopf, A. Busse, and Mc. Angermeyer. Mild cognitive impairment– a review of prevalence , incidence and outcome according to current approaches. *Acta Psychiatr Scand*, 106(6):403–414, 2002.

C.M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer Science+Business Media, LLC, 2006.

F. H. Bouwman, S. N. Schoonenboom, W. M. van der Flier, et al. CSF biomarkers and medial temporal lobe atrophy predict dementia in mild cognitive impairment. *Neurobiology of Aging*, 28:1070–1074, 2007. doi: 10.1002/14651858.CD010386.pub2.

H. Braak and E. Braak. Staging of Alzheimer's disease-related neurofibrillary changes. *Brain Imaging and Behavior.*, 16(3):271–278, 1995. doi: 10.1016/0197-4580(95)00021-6.

E. Cavedo, S. Lista, Z. Khachaturian, P. Aisen, P. Amouyel, K. Herholz, et al. The Road Ahead to Cure Alzheimer's Disease: Development of Biological Markers and Neuroimaging Methods for Prevention Trials Across all Stages and Target Populations. *The Journal of Prevention of Alzheimer's Disease*, 1(3):181–202, 2014. doi: 10.14283/jpad.2014.32.

E. Challis, P. Hurley, L. Serra, M. Bozzali, S. Oliver, and M. Cercignani. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage.*, 112:232–243, 2015. doi: 10.1016/j.neuroimage.2015.02.037.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

R.M. Chapman, M. Mapstone, A.P. Porsteinsson, et al. Diagnosis of Alzheimer's Disease Using Neuropsychological Testing Improved by Multivariate Analyses. *Journal of clinical and experimental neuropsychology*, 32(8):793–808, 2010. doi: 10.1080/13803390903540315.

J. J. Chen, C.-A. Tsai, H. Moon, H. Ahn, J. J. Young, and C.-H. Chen. Decision threshold adjustment in class prediction. *SAR and QSAR in Environmental Research*, 17(3):337–352, 2006. doi: 10.1080/10659360600787700.

B. Cheng, M. Liu, H. Suk, and D. Shen. Multimodal manifold-regularized transfer learning for MCI conversion prediction. *Brain Imaging and Behavior.*, 9(4):913–926, 2015. doi: 10.1007/s11682-015-9356-x.

G. Cornutiu. The Epidemiological Scale of Alzheimer's Disease. *Journal of Clinical Medicine Research*, 7(9):657–666, 2015. doi: 10.14740/jocmr2106w.

D. G. Crenshaw, W. K. Gottschalk, M. W. Lutz, I. Grossman, A. M. Saunders, J. R. Burke, et al. Using genetics to enable Alzheimer's disease prevention studies. *Clinical Pharmacology and Therapeutics*, 93(2):177–185, 2013. doi: 10.1038/clpt.2012.222.

R. Cuingnet, E. Gerardin, J. Tessieras, and G. Auzias. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *NeuroImage.*, 56(2):766–781, 2011. doi: 10.1016/j.neuroimage.2010.06.013.

C. Davatzikos, Y. Fan, X. Wu, D. Shen, and S. M. Resnick. Detection of prodromal Alzheimer's disease via pattern classification of MRI. *Neurobiology of Aging*, 29(4):514–523, 2008. doi: 10.1016/j.neurobiolaging.2006.11.010.

C. Davatzikos, P. Bhatt, LM. Shaw, KN. Batmanghelich, and JQ. Trojanowski. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol Aging.*, 32(12):2322.e19–27, 2011. doi: 10.1016/j.neurobiolaging.2010.05.023.

L. W. De Jong, K. van der Hiele, I. M. Veer, et al. Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. *Brain*, 131(12):3277–3285, 2008. doi: 10.1093/brain/awn278.

T. Fletcher. Support Vector Machines Explained. Technical report, University College London (UCL), 2006. URL http://www.tristanfletcher.co.uk/SVM%20Explained.pdf.

M.F. Folstein, S.E. Folstein, and P.R. McHugh. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *ournal of Psychiatric Research*, 12(3):189–198, 1975.

G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010. doi: 10.1038/nrneurol.2009.215.

J. Golomb, A. Kluger, and S. H. Ferris. Mild cognitive impairment: historical development and summary of research. *Dialogues in Clinical Neuroscience*, 6(4):351–367, 2004.

K. R. Gray, P. Aljabar, RA. Heckemann, A. Hammers, and D. Rueckert. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *Neuroimage*, 65:167–175, 2013. doi: 10.1016/j.neuroimage.2012.09.065.

C. Hinrichs, V. Singh, G. Xu, S. C. Johnson, and the Alzheimers Disease Neuroimaging Initiative. Predictive Markers for AD in a Multi-Modality Framework: An Analysis of MCI Progression in the ADNI Population. *NeuroImage*, 55(2):574–589, 2011. doi: 10.1016/j.neuroimage.2010.10.081.

H. Hippius and G. Neundörfer. The discovery of Alzheimer's disease. *Dialogues in Clinical Neuroscience*, 5(1):101–108, 2003.

C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003. URL http://www.csie.ntu.edu.tw/~cjlin/papers.html.

C. R. Jack, M. S. Albert, D. S. Knopman, G. M. McKhann, R. A. Sperling, M. C. Carrillo, et al. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7(3):257–262, 2011. doi: 10.1016/j.jalz.2011.03.004.

B. Jie, D. Zhang, B. Cheng, and D. Shen. Manifold Regularized Multitask Feature Learning for Multimodality Disease Classification. *Hum. Brain Mapp.*, 36(2):489–507, 2015. doi: 10.1002/hbm.22642.

S. Klöppel, C. M. Stonnington, C. Chu, et al. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(PT 3):681–689, 2008. doi: 10.1093/brain/awm319.

J. Koikkalainen, H. Rhodius-Meester, A. Tolonen, et al. Differential diagnosis of neurodegenerative diseases using structural MRI data. *NeuroImage: Clinical*, 11:435–449, 2016. doi: 10.1016/j.nicl.2016.02.019.

S. Lahmiri and M. Boukadoum. Alzheimer's Disease Detection in Brain Magnetic Resonance Images Using Multiscale Fractal Analysis. *ISRN Radiology*, 627303, 2013. doi: 10.5402/2013/627303.

C.-C. Liu, T. Kanekiyo, H. Xu, and G. Bu. Apolipoprotein E and Alzheimer disease: risk, mechanisms, and therapy. *Nature Reviews. Neurology*, 9(2):106–118, 2013. doi: 10.1038/nrneurol.2012.263.

F. Liu, C. Wee, H. Chen, and D. Shen. Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's Disease and mild cognitive impairment identification. *NeuroImage*, 84:466–475, 2014. doi: 10.1016/j.neuroimage.2013.09.015.

J. Liu and J. Ye. Efficient L1/Lq norm regularization. arXiv:1009.4766v1, 2010.

J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009. Software available at http://www.public.asu.edu/~jye02/Software/SLEP.

L. Liu, L. Fu, X. Zhang, J. Zhang, X. Zhang, B. Xu, et al. Combination of dynamic 11 C-PIB PET and structural MRI improves diagnosis of Alzheimer's disease. *Psychiatry Research: Neuroimaging.*, 233(2):131–140, 2015. doi: 10.1016/j.pscychresns.2015.05.014.

G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and EM. Stadlan. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34(7):939–944, 1984. doi: 10.1212/WNL.34.7.939.

GM. McKhann, DS. Knopman, H. Chertkow, BT. Hyman, CR. Jack, CH. Kawas, et al. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7(3):263–269, 2011. doi: 10.1016/j.jalz.2011.03.005.

L. Meier, S. Geer, and P. Bühlmann. The Group LASSO for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008. doi: 10.1111/j.1467-9868.2007.00627.x.

T.P. Minka. Expectation Propagation for Approximate Bayesian Inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-800-1.

P. Morgado. Image Pre-processing. Technical report, Instituto Superior Técnico, February 2014.

A. D. Murray. *Imaging in Alzheimer Disease and Other Dementias*, volume 22. Neuroimaging Clinics of North America, 2012.

F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, 45(1 Suppl):S199–S209, 2009. doi: 10.1016/j.neuroimage.2008.11.007.

RC. Petersen, RO. Roberts, DS. Knopman, YE. Geda, RH. Cha, VS. Pankratz, et al. Prevalence of mild cognitive impairment is higher in men. *Neurology*, 75(10):889–897, 2010. doi: 10.1212/WNL.0b013e3181f11d85.

A. Rao, Y. Lee, A. Gass, and A. Monsch. Classification of Alzheimer's disease from structural MRI using sparse logistic regression with optional spatial regularization. *In Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 4499–4502, 2011.

C. E. Rasmussen and H. Nickisch. Gaussian Process Regresion and Classification Toolbox version 3.6 for GNU Octave 3.2.x and Matlab 7.x. 2015. Software available at http://www.gaussianprocess.org/gpml/code/matlab/doc/.

C.E. Rasmussen and C.K.I Williams. *Gaussian Processes for Machine Learning*. MIT Press, ISBN 026218253X., 2006.

L. Rosasco, E. De Vito, A. Caponnetto, et al. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004. doi: 10.1162/089976604773135104.

S. Ryali, K. Supekar, D.A. Abrams, and V. Menon. Sparse logistic regression for whole-brain classification of fMRI data. *Neuroimage.*, 51(2):752–764, 2010. doi: 10.1016/j.neuroimage.2010.02.040.

D.P. Salmon and M.W. Bondi. Neuropsychological Assessment of Dementia. *Annual review of psychology*, 60:257–282, 2009. doi: 10.1146/annurev.psych.57.102904.190024.

L. P. Schilling, E. R. Zimmer, M. Shin, A. Leuzy, T. A. Pascoal, et al. Imaging Alzheimer's disease pathophysiology with PET. *Dementia Neuropsychologia*, 10(2):79–90, 2016. doi: 10.1590/S1980-5764-2016DN1002003.

M. Schmidt. *Least squares optimization with L1-norm regularization.* Project Report number CS542B, UBC, University of Alberta, Canada, 2005.

B. Scholkopf and A.J. Smola. *Learning with Kernels.* MIT Press, 2002.

J. L. Shaffer, J. R. Petrella, F. C. Sheldon, K. R. Choudhury, V. D. Calhoun, R. E. Coleman, and P. M. Doraiswamy. Predicting cognitive decline in subjects at risk for Alzheimer Disease by using combined cerebrospinal fluid, MR imaging, and PET Biomarkers. *Radiology*, 266(2), 2013. doi: 10.1148/radiol.12120010/-/DC1.

W.R. Shankle, S. Mani, MJ. Pazzani, and P. Smyth. Detecting very early stages of Dementia from normal aging with Machine Learning methods. *Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine, AIME97*, 1211:73–85, 1997.

RA. Sperling, PS. Aisen, LA. Beckett, DA. Bennett, S. Craft, AM. Fagan, et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7(3):280–292, 2011. doi: 10.1016/j.jalz.2011.03.003.

R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

S. Todd, S. Barr, M. Roberts, and A. P. Passmore. Survival in dementia and predictors of mortality: a review. *International Journal of Geriatric Psychiatry*, 28(11):1109–1124, 2013. ISSN 1099-1166. doi: 10.1002/gps.3946.

K. Uludağ and A. Roebroeck. General overview on the merits of multimodal neuroimaging data fusion. *NeuroImage*, 102, Part 1:3–10, 2014. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2014.05.018.

University of Oxford. Harvard-Oxford cortical and subcortical structural atlases. [online]. URL http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases. [Accessed August, 2016].

P. Vemuri and C. R. Jack. Role of structural MRI in Alzheimer's disease. *Alzheimer's Research Therapy.*, 2(4):23, 2010. doi: 10.1186/alzrt47.

P. Vemuri, H. J. Wiste, S. D. Weigand, L. M. Shaw, J. Q. Trojanowski, et al. MRI and CSF biomarkers in normal, MCI, and AD subjects: Predicting future clinical change. *Neurology.*, 73(4):294–301, 2009. doi: 10.1212/WNL.0b013e3181af79fb.

J.E. Vogt and V. Roth. A complete analysis of the $L_{1,p}$ Group- LASSO. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, page 185–192, Edinburgh, Scotland, UK, 2012.

K. Walhovd, A. Fjell, J. Brewer, L. McEvoy, C. Fennema-Notestine, D. Hagler, et al. Combining MRI, PET and CSF biomarkers in diagnosis and prognosis of Alzheimer's disease. *AJNR. American Journal of Neuroradiology.*, 347(2):1070–1074, 2010. doi: 10.3174/ajnr.A1809.

S. Weintraub, A.H. Wicklund, and D.P. Salmon. The Neuropsychological Profile of Alzheimer Disease. *Cold Spring Harbor Perspectives in Medicine*, 2(4):a006171, 2012. doi: 10.1101/cshperspect.a006171.

C. Williams. Model Selection for Gaussian Processes. [PDF document]. Institute for Adaptive and Neural Computation School of Informatics, University of Edinburgh, UK., December 2006. Retrieved from: http://clopinet.com/isabelle/Projects/NIPS2006/Williams.pdf.

C. K. I. Williams and D. Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998. doi: 10.1109/34.735807.

World Health Organization (WHO) and Alzheimer's Disease International. Dementia: A public health priority. *Geneva: World Health Organization.*, page 92–93, 2012. URL http://www.who.int/mental_health/publications/dementia_report_2012/en/.

J. Ye, T. Wu, and J. Li. Machine Learning Approaches for the Neuroimaging Study of Alzheimer's Disease. *Computer*, 44(4):99–101, 2011. doi: 10.1109/MC.2011.117.

J. Young, M. Modat, M. J. Cardoso, A. Mendelson, D. Cash, S. Ourselin, and the Alzheimer's Disease Neuroimaging Initiative. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2:735–745, 2013. doi: 10.1016/j.nicl.2013.05.004.

L. Yu. Feature selection for high-dimensional data: a fast correlation-based filter solution. In *Proceedings of the twentieth International Conference on Machine Learning (ICML-2003)*, page 856–863, Washington DC, 2003.

M. Yuan and Y. Lin. Model Selection and Estimation in Regression With Grouped Variables. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 68(1):49–67, 2006. doi: 10.1111/j.1467-9868.2005.00532.x.

D. Zhang, Y. Wang, L. Zhou, Shen D. Yuan, H., and Alzheimer's Disease Neuroimaging Initiative. Multimodal Classification of Alzheimer's Disease and Mild Cognitive Impairment. *NeuroImage*, 55(3):856–867, 2011. doi: 10.1016/j.neuroimage.2011.01.008.

D. Zhang, Shen, and Alzheimer's Disease Neuroimaging Initiative. Predicting Future Clinical Changes of MCI Patients Using Longitudinal and Multimodal Biomarkers. *Chen K, ed. PLoSONE.*, 7(3):e33182, 2012. doi: 10.1371/journal.pone.0033182.

S. Zhang, N. Smailagic, C. Hyde, N. Ah, Y. Takwoingi, R. Mcshane, and J. Feng. C-PIB-PET for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews.*, 7(Art.No.:CD010386):131–140, 2014. doi: 10.1002/14651858.CD010386.pub2.

J. Zhou, L. Yuan, J. Liu, and J. Ye. A Multi-task Learning Formulation for Predicting Disease Progression. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 814–822, New York, USA, 2011. ACM. ISBN 978-1-4503-0813-7. doi: 10.1145/2020408.2020549.

X. Zhu, H-I. Suk, Y. Zhu, K-H. Thung, G. Wu, and D. Shen. Multi-view Classification for Identification of Alzheimer's Disease. *Machine learning in medical imaging MLMI*, 9352:255–262, 2015. doi: 10.1007/978-3-319-24888-2_31.