Diagnosis of Alzheimer’s Disease using Multiple Instance Learning

Alexandre Ochôa de Carvalho de Campos Matos
(BSc)

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Prof. Maria Margarida Campos da Silveira
Dr. Durval Campos Costa

Examination Committee

Chairperson: Prof. João Miguel Raposo Sanches
Supervisor: Prof. Maria Margarida Campos da Silveira
Members of the Committee: Prof. Ana Luísa Nobre Fred

June 2015
Nunca conheci ninguém que mudasse tanto como os defuntos, o que me leva a pensar que a morte não é nada do que eu imaginava, o que me leva, até, a questionar a sua existência.

António Lobo Antunes
Acknowledgments

This work was funded by Fundação para a Ciência e Tecnologia (FCT) through the project ADIAR. Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904).

I would like, first of all, to thank my supervisor, professor Margarida Silveira, for her support and helpful ideas during the development of this work. I also express my gratitude to Pedro Morgado, who guided me in many different topics in this thesis and will not forget the various discussions with Filipa Rodrigues.

I am also very grateful towards my family, who has always been present and shaped my way of being. I would be unfair if I did not mention my aunt Doroteia, who very devotedly reviewed this work. Finally, I would like to thank my close friends, and particularly Maria.

I dedicate this thesis to my uncle Carlos and grandfather Fernando, who had not the opportunity to see it finished.
Abstract

Alzheimer’s disease (AD) is a burden for the contemporary society. Its early detection and accurate diagnosis may lead to an improvement in treatment. The diagnosis at pre-clinical stages or the prediction of conversion of patients with Mild Cognitive Impairment (MCI) to AD is a challenging problem, receiving attention because of the immense associated social and economic costs.

This work aims to present an automatic diagnostic tool based on a recent machine learning field, the Multiple-Instance Learning (MIL) model, which differs from the more common supervised learning paradigm.

Concretely, Fluorodeoxyglucose (FDG)-PET images were used as source of data to perform binary classification between AD, MCI and Cognitively Normal (CN) patients. In order to accomplish this, five Multi-Instance Learning algorithms - Diverse Density, CKNN, MILES, YARDS and BARTMIP - were tested, spanning the main MIL paradigms. Experiments were done both in a cross-sectional and longitudinal analysis. Also, an improved and still scarcely used method of normalisation - the Yakushev Normalisation - was implemented, confirming beneficial effects on diagnosis accuracy.

Results show that, although not outperforming other supervised approaches, this paradigm is competitive with the state of the art. Furthermore, no significative differences were found between the five classifiers. Best ballanced accuracies obtained for CN vs. AD, MCI vs. AD and CN vs. MCI were, respectively, 90.98%, 71.32% and 66.85%. On the other hand, Receiver Operator Characteristic (ROC) curves analysis registered areas values of 0.96, 0.77 and 0.67, respectively. What concerns the effect of longitudinal analysis, it seems to be effective only when the follow-up instant shows better accuracy. Finally, it is shown a possible way to overcome image registration through MIL.

Keywords

Alzheimer’s Disease, Computer Aided Diagnosis, Positron Emission Tomography, Multiple Instance Learning
Resumo

A doença de Alzheimer (AD) representa uma calamidade para a sociedade contemporânea. A sua detecção e diagnóstico podem levar a uma melhoria no seu tratamento. O diagnóstico na fase pré-clínica ou a previsão da conversão de pacientes com Décipe Cognitivo Ligeiro (MCI) são problemas desafiadores, recebendo particular atenção pelo seu custo económico e social associado.

Este trabalho pretende apresentar uma ferramenta de diagnóstico automático baseada numa nova área da aprendizagem automática, o modelo Multiple-Instance Learning (MIL), que difere do paradigma supervisionado.

Concretamente, imagens de Flurodeoxiglicose (FDG)-PET foram usadas em classificação binária entre pacientes AD, MCI e cognitivos normais (CN). Para tal, cinco algoritmos de MIL - Diverse Density, CKNN, MILES, YARDS e BARTMIP - foram testados, abrangendo os principais paradigmas de MIL. As análises feitas são transversais e longitudinais. Foi também aplicado um método melhorado e pouco difundido de normalização - Normalização de Yakushev - que confirmou ter efeitos benéficos no diagnóstico.

Os resultados obtidos mostram que, ainda que não ultrapassando abordagens supervisionadas, este paradigma é competitivo com o estado da arte. Além disso, as diferenças entre os cinco classificadores revelaram-se não significativas. As melhores accuracies médias obtidas para CN vs. AD, MCI vs. AD e CN vs. MCI foram, respectivamente, 90.98%, 71.32% e 66.85%. Por outro lado, a análise de curvas de Receiver Operating Characteristic (ROC) obteve áreas de 0.96, 0.77 e 0.66, respectivamente. No que diz respeito à análise longitudinal, esta parece melhorar os resultados apenas quando o instante de tempo de follow-up apresenta melhor accuracy. Finalmente, mostra-se uma possível utilização de MIL como alternativa ao uso de imagens registadas.

Palavras Chave

Doença de Alzheimer, Diagnóstico assistido por computador, Tomografia por Emissão de Positrões, Multiple Instance Learning
List of Figures

1.1 The Amyloid Cascade Hypothesis ................................................. 6
1.2 Regional prevalence of dementia .................................................. 6
1.3 Number of people with dementia in low, middle and high income countries 7
1.5 The Continuum of Alzheimer’s Disease ........................................... 8
1.6 AD prevalence and PiB ................................................................. 10
1.7 Biomarker dynamical model .......................................................... 11

3.1 Supervised Learning and Multiple Instance Learning ............................ 24
3.2 The ambiguity spectrum ............................................................... 25
3.3 An illustration of bags and instances in a bidimensional space ................ 29
3.4 Learning under the Standard Multi-Instance (SMI) assumption ................. 30
3.5 Weidmann’s hierarchy ................................................................. 31
3.6 Instance Space ............................................................................. 32
3.7 Bag space paradigm ...................................................................... 34
3.8 Embedded space paradigm ............................................................. 35

4.1 Image registration steps ............................................................... 39
4.2 Reference cluster ....................................................................... 40
4.3 The binary mask ........................................................................ 41
4.4 Patch selection ........................................................................... 42
4.5 The maximum optimal margin ....................................................... 43
4.6 Non-linearly separable data .......................................................... 45
4.7 Illustration of Hausdorff distance .................................................. 48
4.8 Majority vote errors in CKNN ....................................................... 49
4.9 Citers and references ................................................................... 49
4.10 Diverse Density .......................................................................... 51
4.11 Diverse Density optimisation ....................................................... 53
4.12 DD single point-and-scaling concept ............................................. 54
4.13 MILES embedding .................................................................... 55
4.14 MILES pseudocode ................................................................... 57
4.15 YARDS pseudocode ................................................................... 59
4.16 BAMIC pseudocode ................................................................... 60
# List of Tables

2.1 Performance of different multiple-instance learning CAD systems applied to Alzheimer’s Disease .......................................................... 20

2.2 Performance of different longitudinal CAD systems applied to Alzheimer’s Disease .......................................................... 20

2.3 Performance of different CAD systems applied to Alzheimer’s Disease .......................................................... 21

3.1 Accuracy of benchmark MIL algorithms in Musk and COREL data sets .......................................................... 28

4.1 Reference cluster best parameters ........................................................................ 40

4.2 Patch selection parameters ........................................................................ 42

5.1 Data set clinical and demographic information .......................................................... 68

5.2 P-values for regional voxel intensity average differences between clinical groups .......................................................... 70

5.3 Most discriminative regions by maximum average t-values .......................................................... 71

5.4 Best and worst patch parameter combination for SVM with patches at the baseline .......................................................... 74

5.5 Best and worst patch parameter combination for SVM with patches at 12 months .......................................................... 74

5.6 Best and worst patch parameter combination for CKNN at the baseline .......................................................... 75

5.7 Best and worst patch parameter combination for CKNN at 12 months .......................................................... 75

5.8 Best and worst patch parameter combination for DD at the baseline .......................................................... 76

5.9 Best and worst patch parameter combination for DD at 12 months .......................................................... 76

5.10 Best and worst patch parameter combination for MILES at the baseline .......................................................... 77

5.11 Best and worst patch parameter combination for MILES at 12 months .......................................................... 77

5.12 Best and worst patch parameter combination for YARDS at the baseline .......................................................... 79

5.13 Best and worst patch parameter combination for YARDS at 12 months .......................................................... 80

5.14 Best and worst patch parameter combination for BARTMIP at the baseline .......................................................... 81

5.15 Best and worst patch parameter combination for BARTMIP at 12 months .......................................................... 81

5.16 Best and worst patch parameter combination for SVM with patches (longitudinal analysis) .......................................................... 83

5.17 Best and worst patch parameter combination for MILES (longitudinal analysis) .......................................................... 83

5.18 Best results for each classifier (CN vs. AD) .......................................................... 85

5.19 Best results for each classifier (CN vs. MCI) .......................................................... 85

5.20 Best results for each classifier (MCI vs. AD) .......................................................... 86
Acronyms

Aβ β Amyloid

ACH Amyloid Cascade Hypothesis

AC-PC Anterior Commissure - Posterior Comissure

ADNI Alzheimer’s Disease Neuroimaging Initiative

AD Alzheimer’s Disease

APR Axis Parallel Rectangle

BARTMIP Bag-level Representation transformation for Multi-Instance Prediction

APOE Apo lipoprotein E

APP Amyloid Precursor Protein

AUC Area Under the Curve

BS Bag Space

CAD Computer Aided Diagnosis

CDR Clinical Dementia Rating

CGM Cerebral Global Mean

CKNN Citation K-nearest neighbour

CMR Cerebral Metabolic Rate

CN Cognitively Normal

CSF Cerebrospinal Fluid

CT Computed Tomography

CV Cross-Validation

DD Diverse Density

DICOM Digital Imaging and Communications in Medicine
DSM  Diagnostic and Statistical Manual of Mental Disorders

ES  Embedded Space

FAD  Familial Alzheimer’s Disease

18FDG  18F-2-fluoro-2-deoxy-D-glucose

GM  Gray Matter

IS  Instance Space

LDA  Linear Discriminant Analysis

MCI  Mild Cognitive Impairment

MIL  Multiple-Instance Learning

MILES  Multiple-Instance Learning via Embedded Instance Selection

MNI  Montreal Neurological Insitute and Hospital

MMSE  Mini Mental State Examination

MRI  Magnetic Resonance Imaging

NFT  Neurofibrillary Tangle

PAC  Probably Approximately Correct

PAD  Preclinical Alzheimer’s Disease

PCA  Principal Component Analysis

PET  Positron Emission Tomography

PiB  Pittsburgh compound B

PS1  Presenilin 1

PS2  Presenilin 2

QDA  Quadratic Discriminant Analysis

ROI  Region Of Interest

SAD  Sporadic Alzheimer’s Disease

SDAT  Senile Dementia of Alzheimer’s Type

SMI  Standard Multi-Instance

SPM  Statistical Parametric Mapping

SPECT  Single-Photon Emission Computed Tomography
SVM  Support Vector Machines

TI-ETDS  Time Independent Eye Track Driven Selection

TD-ETDS  Time Dependent Eye Track Driven Selection

VC-dimension  Vapnik Chervonenkis dimension

VI  Voxel Intensity

YARDS  Yet Another Radial Distance-based Similarity measure
Introduction

Contents

1.1 Motivation ................................................................. 2
1.2 Proposed Approach ..................................................... 11
1.3 Original Contributions ................................................. 12
1.4 Thesis Outline ............................................................ 12
1.1 Motivation

1.1.1 Dementia - the history of a burden

Dementia is not a recent topic in scientific debate. In fact, the first writings referring to it appear in Ancient Greece, dating from the 7th century BC [1]. At that time, philosophers understood dementia as an inevitable feature of ageing, related to a disruption in the equilibrium of body humours. Pythagoras (571 BC - 495 BC) considered the age above 63 as what he called senium, a period of decline and decay of the human body and regression of mental capacities. Later, the roman Galen (130-201) introduced the term morosis to describe dementia, including advanced age as one of the situations when it might happen. As he described, those suffering from morosis were “people in whom the knowledge of letters and other arts are totally obliterated; indeed they can’t even remember their own names” [2].

During the Middle Ages, little was done on mental health. However, for the first time a philosopher associates the brain with the source of memory. Indeed, in a treaty entitled Methods of Preventing the Appearance of Senility, Roger Bacon (1214-1294) states that “in the posterior part [of the brain] occurs oblivion (...) [and] old age is the home of forgetfulness” [2], a notable remark, since until then the heart was considered the source of thought [1].

Then, in the 17th century, Thomas Willis (1621-1675), an anatomist, explores the causes of morosis, splitting them in six categories: congenital factors, age, alcohol and drug abuse, head injury, disease and epilepsy. By then, the concept of dementia comprised a generic state of intellectual deficit, whatever the cause and age, and various synonymous designations coexisted such as lethargy, stupidity or insensitivity [3]. Only in the nineteenth century, however, would mental disorders, and particularly dementia, receive due attention from the medical community. The french Pinel (1745-1826) and his student Esquirol (1772-1840) both contributed to the development of a deeper understanding of this subject, introducing systematic clinical observation and a more precise terminology in psychiatry. The latter further described the stages of cognitive decline. Most importantly, Bénédict Morel (1809-1873), in his Traité des Maladies Mentales, notices for the first time that brain weight decreases under dementia, coining this weight loss as a constant feature of dementia. In 1864, Wilks (1824-1911) defines brain atrophy, one of the major features of dementia. His words were “Instead of the sulci meeting, they are widely separated and their intervals filled with serum and which, on being removed with the pia mater, the full depth of the sulci can be seen” [1].

General paresis was the first dementing disease to be clearly described. It was noticed by the medical community that cerebral vasculature was affected by this condition, which called attention towards cerebrovascular changes. Soon dementia started to be seen as a result of diminished blood supply to the brain. At the same time, with new histochemical techniques, it became possible to better understand the etiology of dementia. Blocq (1860-1896) and Marinesco (1863-1938) first described the accumulation of an unknown substance into plaques in a patient with epilepsy in 1882, a feature which remained unnoticed [3], until later, Alois Alzheimer (1864-1915), observing the neurons of a deceased patient, could perceive tangles of fibrils within the cytoplasm, but also the presence of plaques of a mysterious substance scattered through the entire cortex, especially in the upper layers. The patient
was a middle-age woman, which did not fit into senile dementia, but rather into pre-senile dementia [4]. The clinical and neuropathological presentation of his patient not fitting any of the known cases of dementia led Alzheimer to suggest another type of dementia, which would be baptized with the name of Alzheimer’s Disease (AD) by Emil Kraepelin (1856-1926), when publishing the eighth edition of *Textbook of Psychiatry* in 1910. In the same book, Kraepelin described *arteriosclerotic insanity* based on previous work done by Alzheimer and Binswanger (1852-1929), a concept that would evolve to vascular dementia [5] after work by Hachinsky et al [6].

In the following decades a deep debate was held around the origin of tangles and plaques, but also on the possibility of genetic causes of AD. Moreover, the distinction between senile dementia and AD became a hot topic. The majority of patients reported in Europe were under the age of 65, which precluded them from being cases of senile dementia. In fact, by then, disease was essentially classified based on age and the emphasis given to it was particularly felt in the field of mental diseases, where the traditional negative view about old age prevailed [7]. Additionally, since dementia was regarded as a natural degenerescence in the elder person, medical treatment was seldom sought [4].

In 1927, Divry identified the plaque substance as amyloid, which would only be confirmed in the 1960s with electron microscopy. What may impress the reader is that in the 1950s, there was still the erroneous idea that AD was a rare disorder [8]. Eventually, it was concluded that the only difference between AD and senile dementia was the onset of the symptoms. Wherefore, they were unified under the name of Senile Dementia of Alzheimer’s Type (SDAT), a term no longer used nowadays.

In the 1990s research led to a boost in the comprehension of the genetics behind AD, with the discovery of three genes: the APP in 1991, the presenilin 1 in 1995 and presenilin 2 in 1996 [9]. All these were involved in the formulation of one of the most influential theories behind the causes of AD, known as the Amyloid Cascade Hypothesis (ACH) [10], defended by the so called baptists. Soon, another train of thought appeared, led by tauists [9], [11].

By the end of the 20th century, an incredible range of new technologies, mainly in the medical imaging area, allowed a deeper and easier assessment of dementia, namely Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI). New concepts, such as Mild Cognitive Impairment (MCI) (1988), arose, helping to identify early features of dementia [12]. Nonetheless, many hazy questions remain to be answered.

### 1.1.2 Alzheimer’s Pathology

AD is seen nowadays as a neurological disorder which affects mostly people aged 65 or over. It belongs to a wider category of diseases all termed as dementia, such as vascular dementia (20-30%), frontotemporal lobar degeneration (5-10%), dementia with Lewy bodies (5 %) and the less studied mixed dementia and Parkinson’s disease. AD accounts for 50-75% of the cases [13].

Dementia, in broad terms, is a syndrome related to disease in the brain, with progressive decline in cognitive abilities. Although it can appear at any age, it is more common over the age of 65, with its prevalence doubling every 5 years after this cornerstone and its numbers doubling every twenty years [13]. The diagnosis of dementia is usually carried out according to the Diagnostic and Statistical
Manual of Mental Disorders (DSM) criteria, which require a decline in memory and a simultaneous impairment in one cognitive function such as ability to speak coherently and understand spoken or written language, ability to recognize or identify objects, ability to perform motor activities or ability to think abstractly. All these must interfere with daily life [14].

In terms of its symptoms, AD gradually worsens the ability to remember new information. Soon, changes in mood and personality start to appear, with apathy and depression often affecting the patient. In middle and final stages there is time or place notion loss, difficulty in planning daily tasks, decreased judgement, withdrawal from work or social activities and trouble with speaking and writing [14].

What concerns the AD diagnosis, this is based on the criteria published in 1984 by the Alzheimer’s Association and the National Institute of Neurological Disorders and Stroke [15]. These have shown a sensitivity of 81% and specificity of 70% [16]. However, many developments have taken place since then. In particular, other dementia syndroms such as vascular dementia and dementia with Lewy bodies were not well described by then. Moreover, imaging techniques were not taken into account at the time and genetics of AD was poorly understood [16]. All these factors led to the release of new criteria, published by the National Institute on Aging (NIA) and the Alzheimer’s Association in 2011. These update the ones published before, introducing new and relevant changes: they identify three stages of AD - Preclinical Alzheimer’s Disease (PAD); MCI; dementia due to AD - and launch biomarker tests. The latter are based on biomarkers, that is, biological factors that can indicate the presence of a disease [17].

A word is necessary on the three stages proposed. In fact, they allow to disclosure the beginning of the disease without having yet the symptoms, in its preclinical stage [18]. The first one, PAD, already presents measurable changes in the brain, cerebrospinal fluid and blood, but no symptoms. This stage reflects the possibility, suggested by some researchers, that Alzheimer’s Disease starts up to 20 years before symptoms arise [19, 20]. Regarding the MCI stage, this presents the first noticeable changes in thinking, despite not affecting daily activities. Some patients in this stage show concern about some changes in cognition and there is a lower performance in one or more cognitive domains such as memory, language or visuospatial skills, greater than the normal for their age, while maintaining their independence when performing daily tasks [21]. Finally, the probable AD patient shows interference with the ability to do usual activities, a decline in performance and cognitive impairment, which must be assessed through a history-taking from the patient and an objective cognitive exam, and encompasses at least two behavioural domains amongst memory, judgement, visuospatial abilities, language and personality. Exams such as Mini Mental State Examination (MMSE) and Clinical Dementia Rating (CDR) are often used to assess the patient mental status. Despite all the data collected from the patient, it should be emphasised that AD definite diagnosis is declared post-mortem.

With respect to the risk factors, a lot of work has still to be done, but it is generally accepted that age is the main one. In early onset cases, before age 65, genetic causes seem to play an important role. Three autosomic genes have been identified - Amyloid Precursor Protein (APP), Presenilin 1 (PS1) and Presenilin 2 (PS2) - all of them associated to an increase in amyloid production and autosomic dominant [22]. Early onset accounts for about 2% of the cases of the disease, but 60% of these are
Familial Alzheimer’s Disease (FAD) (two or more people have been diagnosed in the family) against Sporadic Alzheimer’s Disease (SAD). In the general affected population, the familial form of AD accounts for 25% of the numbers. The Apo lipoprotein E (APOE) gene seems to have some weight of its ε4 form, and is the major genetic risk for late onset AD [18]. A hallmark case which shows evidence of genes is Down Syndrome, in which more than half of individuals develop AD after the age of 40 due to the overexpression of the APP gene present in the 21th chromosome [23]. Besides, studies of twins suggest that environmental factors are relevant. Other agents considered to increase the risk are smoking [24], head injury, limited education, diabetes, hypertension and atherosclerosis [25].

Among the various theories trying to explain the origin of the disease, two assume particular importance - the Amyloid Cascade Hypothesis (ACH) and the tau hypothesis. The first proposes that the deposition of β Amyloid (Aβ) is the first pathologic event, leading to ensuing phenomena such as the formation of Neurofibrillary Tangles (NFTs) and cell death [10]. Aβ peptides are the major components of senile plaques in AD, but also in dementia with Lewy bodies and Parkinson’s disease dementia. Their origin lies on the Amyloid Precursor Protein (APP) when cleaved by the enzymes γ and β-secretase [26]. Senile plaques form inside neurons and some experiments show that high levels of APP indeed lead to Aβ deposition (mainly the Aβ42 variant) outside the neurons, synaptic loss and gliosis in mice. However, some evidence leads some of the literature to consider Aβ as a reactive feature rather than causal, that is, consequence of other process, and not a cause for the subsequent events in the disease. In fact, APP production seems to be a result of some form of head injury and moreover it shows signs of helping to maintain cell function [27]. Besides, ACH considers tau production as a downstream event, as consequence of Aβ deposition, which is not corroborated by both the spatial and temporal pattern of the latter and tau deposition [11] (see figure (1.1)).

These and other lacunae in ACH contributed to the rise of the tau hypothesis. This states that microtubules of associated tau protein become highly phosphorylated and aggregated in tangles when tau is impaired, which leads to accumulation inside cells and ultimately to neuronal death. Indeed, there is a good correlation between the number of tangles and cognitive function [28]. Furthermore, tauopathies are involved in corticobasal degeneration and frontotemporal dementias and are associated with the severity of dementia [29]. Braak et al. [30] divided tau pathology in six stages (I-VI) according to its extent and anatomic presence. Tau protein starts by accumulating in the entorhinal cortex, then hippocampus and finally the neocortex [31], while amyloid plaques are most abundant in frontal cortex, cingulate gyrus, precuneus and lateral parietal and temporal regions [32].

Both hypothesis could merge into one if the mechanism linking Aβ to changes was known. Some evidence leads to the belief that the Wnt pathway may play a role as the unifier of both theories and presents a possible future therapeutic having as target this pathway [33].

As to the macroscopic aspect, AD is characterized by a manifest brain atrophy, mainly at cortical level, hippocampus, entorhinal cortex and mesial temporal regions. Besides, there is a reduction in brain metabolism. A recent study brought interesting results concerning the anatomic differences between normal ageing and AD atrophy. Beyond the previous knowledge that normal ageing atrophy
is smaller, this study shows that the pattern is localized in different regions, even if some are common to both processes [34].

![Diagram of the Amyloid Cascade Hypothesis](image)

**Figure 1.1:** The Amyloid Cascade Hypothesis. Source: [10]

### 1.1.3 Alzheimer’s Disease and its impact on society

According to most recent data from Alzheimer’s Disease International, there are about 44 million people worldwide living with dementia, number expected to double every 20 years and grow up to 135 million in 2050. Regional age specific prevalence ranges from 5% to 9%, being fairly similar (see figure 1.2). Despite the fact that high income countries have a greater percentage of elderly people, only 38% of people with dementia are estimated to live in those countries, a value that is likely to decrease, as population growth boosts in low and middle income countries [35] (see figure 1.3). In terms of gender, more women have dementia (figure 1.4), a fact that appears to result only from a greater life expectancy, age being one of the most important factors in the development of dementia [36].

![Age-specific prevalence of dementia discriminated by world regions, showing previous predictions (2009) and current ones (2013)](image)

**Figure 1.2:** Age-specific prevalence of dementia discriminated by world regions, showing previous predictions (2009) and current ones (2013). Source: [35]

The numbers above make dementia a global epidemic. Its global cost to society was estimated to be US$ 604 billion - 1% of global GDP [37]. Furthermore, a disturbing fact is that there is more funding to cancer and heart disease research, even though dementia represents a higher cost to society, primarily materialised in care homes and informal care given by families. Data from the U.S reveal an impressive reality: more than 15 million americans are caregivers to people with AD. From these, the majority are women. Caregivers usually experience some sort of physical and emotional stress [36].
Up to the moment, no treatment slows or cures the disease, although some can improve symptoms. Some studies have been made with non-steroidal anti-inflammatory drugs (NSAIDs), hormone replacement therapy (HRT) and cholesterol lowering therapies (statins). Besides, antihypertensive treatment also appears to be ineffective [38]. Life expectancy of people with dementia is about 7 years after the diagnosis, although some can live up to 20 years beyond. Most of these years belong to the most severe state of AD, which greatly contributes to the public health impact of this disease [36].

1.1.4 Biomarkers and the role of PET

Biomarkers play increasingly a more relevant role in research and diagnosis of diseases. According to the World Health Organization, they are defined as any measurement reflecting an interaction between a biological system and a potential hazard, which may be chemical, physical, or biological. The measured response may be functional and physiological, biochemical at the cellular level, or a molecular interaction [40]. This includes trivial measures as pulse and blood pressure [41].

AD is not an exception, and various biomarkers exist that allow the improvement of its diagnosis. A testifying fact is the creation of Alzheimer’s Disease Neuroimaging Initiative (ADNI) in 2004, which has as one of its fundamental objectives the identification and validation of new biomarkers for AD [42].
In this context, Mueller et al. [22] and Weiner et al. [43] state that the ideal biomarker should be able to detect a fundamental feature of AD pathology, be minimally invasive, with a simple analysis and not expensive.

AD biomarkers are relevant since cognitive measures have less statistical power and higher variability. Withal, they can act as predictors of cognitive decline and measure the changes that occur in the so-called continuum of AD [43] (see figure (1.7)). That said, they have potential to be used as a clinical and prognostic tool, eventually allowing the effects of treatments to be measured. However, there are still some gaps to be filled, such as biomarkers that can reliably distinguish between different types of dementia [42].

Although no strict nomenclature exists, biomarkers in dementia research can be divided into genetic, biochemical, electrophysiologic and imaging biomarkers [22, 44]. Regarding the genetic ones, the gene APOE-ε4 appears to be the most reliable tool, although its use as diagnostic is limited due to the lack of knowledge and complexity of causes of the disease. With respect to biochemicals, many have been studied, both specific for AD (Aβ, Cerebrospinal Fluid (CSF) tau protein) and non-specific (inflammation and oxidative stress biomarkers). Nonetheless, some limitations and doubts still exist around many of them. The latter type, imaging data, is an area of flourishing new procedures in which there are the structural - CT and MRI - and the functional markers - PET and Single-Photon Emission Computed Tomography (SPECT). MRI has a number of advantages compared to CT, since it allows higher resolution, good gray-white matter discrimination and no bone hardening in the temporal lobe region [45]. Conversely, functional markers are better suited to early stages of the disease since they measure neuronal energy metabolism, which comes across beforehand. PET has a better spatial resolution and correlates better with the degree of disease.

PET is a paradigmatic case of functional markers and a nuclear medicine imaging technique that allows non-invasive imaging of body physiology. Its principle lies on the emission of gamma rays released from a radionuclide - known as tracer or radiotracer - which is introduced in the body incorporated in a biological active molecule.

In particular, 18F-2-fluoro-2-deoxy-D-glucose (18FDG) PET is currently one of the most used techniques. While carbon and oxygen tracers have a small half-life (2 and 20 minutes, respectively), fluorine’s half-life is sufficiently long (110 minutes) [46], which makes it suitable for clinical studies. In brain studies it is found to be very useful to measure glucose metabolism. Although the brain only
accounts for 2% of the body weight, its glucose consumption is approximately 60% [47] and 20% in the case of oxygen, which is used in glucose oxidation to provide energy required for synaptic action potentials. Thus, PET imaging using $^{18}$FDG enables a quantitative estimate of local cerebral metabolic rate of glucose consumption (CMR$_{glc}$). In healthy subjects, the variance of CMR is smaller than in people with dementia, in whom there is a high inter-regional metabolic variability [48]. Most importantly, PET studies have shown that there is a reduction in CMR in people with dementia [49–51]. This also happens with normal ageing, but to a lesser extent. Areas most affected tend to be neocortical association areas, such as posterior cingulate and temporoparietal and frontal multimodal association cortex [52,53]. In AD, the basal ganglia, primary motor and visual cortex and cerebellum are preserved, contrasting with other dementia types [46].

Aside from these desirable features, Silverman et al. [54] have demonstrated that PET undoubtedly improves accuracy when compared to conventional clinical diagnosis, as confirmed by histological analyses, particularly in early AD cases.

Some new and promising techniques have been developed in the past few years. A MRI-based tool, diffusion tensor imaging (DTI), has allowed a new view on the role of white matter in AD progression. DTI measures water diffusion at a molecular level and enables a reliable visualisation of abnormalities in its pattern. Some interesting results have shown differences between normal ageing and subjects with dementia regarding the white matter, which seems to be most affected in the posterior brain, in structures such as the superior longitudinal fasciculus, the cingulum bundle and the corpus callosum [55]. Also, magnetic resonance spectroscopy has seen deep developments. Some of the metabolites currently being studied are N-acetyl-aspartate (NAA), choline (Cho), myo Inositol (mI) and creatine (Cr). NAA, produced in mitochondria, is now considered a neuronal marker, being correlated with the amount of plaque and tangle fibrils. Subjects with AD reveal a net reduction in parietal, temporal and frontal lobe. Simultaneously, there is an increase in mI levels. What concerns choline, studies report ambiguous results, and creatine value is used as normalisation factor for values of other molecules because of its constant value along time. With regard to the use of these biomarkers in differential diagnosis, it is known that many of them overlap in different dementias such as AD and vascular dementia. However, the location of the metabolite change may distinguish the disease. For instance, the NAA levels are more reduced in white matter of patients with vascular dementia [56].

Another important biomarker is Pittsburgh compound B (PiB), a derivative of thioflavin T (an amyloid die) which is used with PET. Many studies show a higher retention of this molecule in AD subjects and inverse correlation with glucose hypometabolism in some areas. Moreover, it has been confirmed that amyloid plaque accumulation is slow and reaches a plateau in MCI, which suggests the latent period of time until the first symptoms of AD arise [32]. Despite these positive findings, some drawbacks do exist, in particular the fact that the number of PiB positive asymptomatic healthy people increases with age, which represents a decrease in the specificity with age. On the other hand, a recent study carried by AIBL [20] showed that APOE-ε4 gene allele carriers are more likely to have high PiB binding. Besides, it demonstrated that amyloid deposition is almost inevitable with age in healthy controls. Remarkably, it was inferred that the prevalence of AD follows the same behaviour.
along time as the prevalence of amyloid deposition in healthy controls with a lag of around 15 years (see figure 1.6). Consistent with this fact, another study asserts that PiB may detect the pathologic changes of AD before the clinical evidence [57].

Figure 1.6: Comparison of the prevalence of PiB binding (blue dots) with beta amyloid deposition in cognitively normal subjects as detected at post-mortem (green triangles) and with the age prevalence of AD in general population (red diamonds). The shape of the data is similar, suggesting that AD only manifests some 15 years after the first structural or functional signs. Source: [20].

Great attention has been given recently to blood-based markers as an alternative to the ones presented above, which are either expensive or invasive. Concerning the widely studied amyloid pathophysiology, the ratio of Aβ42 and Aβ40 in blood has been shown to be clinically meaningful, despite the heterogeneity in results [58]. Another paper refers the discovery of 10 blood-borne markers that could distinguish between AD and controls with 96% sensitivity and 92.5% specificity [59].

Some of the difficulties to be faced by blood biomarkers is the existence of the blood-brain barrier, which prevents some molecules from entering and exiting the brain, although this barrier has a loss of integrity in AD, or the blood proteome, which is highly complex, changing in response to various factors. Also, there is a lack of standardisation methodology. These and other factors imply that there is no fully validated blood-based marker of AD at the moment [60]. In order to circumvent these difficulties, a consortium has been created by the Alzheimer’s Association in 2010, bringing together academic researchers and industry partners [61]. Overall, there is a wide range of potential biomarkers to be used in the future, with promising implications in Computer Aided Diagnosis (CAD) of AD.
1.2 Proposed Approach

The scientific community has shown increasing interest towards the prediction of AD and MCI due to its impact in diagnosis and treatment of this disease. This thesis aims to develop a reliable automatic diagnostic tool against normal controls in the context of Multiple-Instance Learning, a flourishing new area in machine learning.

The data used in this thesis was retrieved from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database [22]. In particular, PET data was analysed after some preprocessing steps. One of the challenges that often arise in machine learning data is the high dimensionality of input images, a concern known as curse of dimensionality, which jeopardises classification results. In order to circumvent this, a feature selection step is usually carried out. In this thesis, a t-value based feature selection method through image patches was tested. In terms of feature extraction, Voxel Intensity (VI) was used. In addition, a normalisation step known as Yakushev Normalisation [62] was applied, showing improved results against the usual cerebral global mean.

Most importantly, regarding classification, five MIL algorithms belonging to three different sub-paradigms - Instance-space, Bag-space and Vocabulary based - were applied along with the renowned supervised Support Vector Machines (SVM) in order to compare Multiple-Instance Learning (MIL) and one of the hallmarks of supervised learning. Concretely, the five MIL algorithms were Diverse Density (DD), Citation K-nearest neighbour (CKNN), Multiple-Instance Learning via Embedded Instance Selection (MILES), Yet Another Radial Distance-based Similarity measure (YARDS) and Bag-level Representation transformation for Multi-Instance Prediction (BARTMIP). A Cross-Validation (CV) procedure was used to assess the methods general performance. In addition, a longitudinal evaluation using SVM and MILES was performed, which allowed a comparison with cross-sectional analysis in the baseline and 12 months data, and a simulation of non-registered images classification was performed with MILES.
1.3 Original Contributions

This thesis represents a new approach to the diagnosis of AD. While not exploring new ways of feature extraction nor selection, which were explored by precedent works such as Morgado [63] and Bicacro [64], among others, MIL algorithms never before experimented in CAD of AD are now applied in this field, which, to the best of our knowledge, was only done by Tong et al. [65]. This represents an innovative way of looking into this disease. In fact, research in the area has been mainly focused in supervised learning methods, neglecting the potential of treating the brain as an entity whose parts have unknown labels, as MIL does. Consequently, MIL is shown to be an alternative to supervised learning.

Besides, an innovative normalization step, the Yakushev Normalisation, still not widely applied in the literature, is applied in this work, showing improved accuracy, sensitivity and specificity. Finally, a simulation of classification with non-registered images is performed, showing some promising results.

1.4 Thesis Outline

The remainder of this thesis is structured as follows. Chapter 2 addresses the state of the art focusing first on the enhancing Yakushev normalisation applied on this work. Then, feature selection and extraction are also addressed in a quick glance due to its importance in CAD field. In chapter 3, a thorough overview of MIL is given, with the presentation of the main paradigms and basic mathematical concepts pertaining to this field. After that, in chapter 4, there is a minute description of the methodology and classification procedure, with a focus on the explanation of the algorithms and their underlying ideas. Chapter 5 is devoted to the presentation of results and finally, chapter 6 is dedicated to the discussion of the results and to the presentation of some suggestions for future work.
2.1 Introduction

Computer aided diagnosis appeared in the 1960’s as an alternative approach to automated computed diagnosis. While the latter intends to replace medical function, the main objective of CAD is to allow an improved diagnosis together with medical evaluation [66]. Since then, systematic research has been developed in areas such as cardiovascular diseases, lung cancer and breast cancer.

CAD is usually structured in three stages: 1) candidate generation, in which suspicious or meaningful traits in images are identified; 2) image feature selection/extraction to characterise the suspicious hallmarks; 3) finally, decision making, in which the diagnosis is established by the doctor with the support of the classifier output [67].

Machine learning algorithms are particularly useful in the last step of CAD, and a variety of them has been applied in the medical field, mainly in a supervised way, such as SVM or neural networks, under which the machine is told the label of training examples and learns a hypothesis that ideally has no error when classifying unseen examples. Multiple-Instance Learning (MIL), a novel approach inside machine learning, augurs new perspectives and perhaps better results for diagnosis of diseases where there is a certain ambiguity, such as Alzheimer. In fact, some areas of the brain may be affected, showing signs of the disease, and simultaneously other areas share traits with healthy subjects.

In the last two decades, many reliable CAD systems have been developed allowing to distinguish AD patients from healthy individuals, and to a lesser extent between AD vs. MCI and Cognitively Normal (CN) vs. MCI. Although no evolution pattern of the disease is well established, its early diagnosis is paramount to improve the efficacy of new treatments.

This chapter is structured in the following way: firstly, Yakushev Normalisation is presented, emphasising its benefits in the context of AD image analysis. Secondly, in section 2.3, the importance of feature selection and extraction is highlighted with an overview of some methods applied to Alzheimer’s Disease. Finally, some work in CAD systems relating to AD, both inside MIL and supervised learning, and considered most valuable to the current thesis, is introduced.

2.2 Yakushev Normalisation

Recently, an important enhancement in diagnosis accuracy was accomplished by Yakushev et al. with respect to reference region normalisation in statistical parametric mapping [68]. Normalisation of regional tracer uptake is a usual proceeding in PET analyses, since it allows to account for inter and intra subject variability. Another benefit lies on the reduced variability of relative CMRglc in comparison to its absolute values. Still, this step is problematic for the widespread metabolic and blood flow reduction. The net effects are an underestimation of metabolic reduction in diseased brains and also the detection of false hypermetabolic regions in AD patients. This can easily be understood when a Cerebral Global Mean (CGM) normalisation is carried through, as in AD brains the CMRglc are abnormally low, which leads to a diminished CGM and to the aforementioned grievous effects. These can appear even in early cases of MCI [69]. The apparently hypermetabolic regions, as Yakushev denominated them, are areas usually spared in AD, such as the primary sensorimotor cortex.
cerebellum, basal ganglia and brain stem. In order to circumvent this weakness, strategies such as
normalising to one of these regions were applied [70]. Others have applied a ROI based approach [71].
Costa [72] and Schmidt [73] analysed the same problem but in the context of cerebral brain flow (CBF)
measure in SPECT. The latter concluded that when global CBF (gCBF) decreases more than regional
CBF (rCBF), normalisation to gCBF detects a false increase in this region, which translates into an
increased probability of detecting false negatives (type II error). Analogously, when there is a greater
increase in gCBF than in rCBF, the region will evince a decrease and there is a bigger probability of
false positive detection (type I error). In short, Schmidt emphasises the need to have the reference
region measure constant whenever one wants to extract true information in a regional analysis.

In a first paper, Yakushev et al. [62] compared the performance of CGM, cerebellar and sensorimo-
tor cortex normalisation, concluding that the latter presented the best results. CGM only accounted
for one third of extent of metabolic deficits in moderate and severe patients, failing to detect areas of
hypometabolism. This proves that the until then usual assumption that regional metabolic impair-
ments do not affect global cerebral metabolism is incorrect. The most notorious effect of applying a
regional reference area was an increase in sensitivity.

In their second paper, Yakushev et al. [68] tested the normalisation to a reference region named
reference cluster. This region is obtained through a voxelwise t-test with AD patients and controls
and then a selection of the voxels with t-value above a given threshold. By doing so, the best of the
areas which are apparently hypermetabolic is chosen. Borghammer et al. [74] called this method the
Yakushev Normalisation and tested it, obtaining similar results. The authors stated that the t-value
threshold may present a caveat and advise not to use a value above 2, since this can lead to false positive
decreases. Although not referring to Yakushev, Chen et al. [75] applied a similar normalisation based
on a voxel cluster termed spared region-of-interest (sROI), selected through a t-score threshold. Other
authors, such as Rodrigues and Silveira [76] and Gray [77], have exploited the benefits of Yakushev
Normalisation.

2.3 Feature Extraction and Selection

Feature selection is a key step in many classification problems. Often one is faced with problems
where there are few subjects and much more variables or features, which may result in poorer classi-
fication performance. Some of the benefits of feature selection are a better visualisation of data, with
a consequent better understanding of the problem, the reduction of storage requirements and also of
training and testing time. Finally, and maybe the most important aspect, it allows the avoidance of
the so called curse of dimensionality, which frequently jeopardises the classifier performance [78]. As
Bishop notes, high dimensionality spaces become sparse. This can easily be seen by computing the
fraction of volume of a thin shell of width $\epsilon$ near the surface of a sphere in a high dimension space.
For a 20 dimension space, a sphere of unitary radius will have almost all its volume in a surface shell
with $\epsilon = 0.2$ [79]. The high dimensionality of input space when there is a relatively small sample of
subjects represents a drawback as it leads to increased complexity and overfitting, that is, the classifier
fits very well the training data but has a poor generalisation power.
With regard to disease diagnosis, features should be representative of the disease process and be robust, that is, the most insensitive as possible to noise or other artifacts [80]. Particularly, when dealing with brain image data, the number of features can easily reach hundreds of thousands, while data sets have usually under 200 subjects and not all image information is useful for classification. This is the case of the present work (see chapter 5). Some common approaches to overcome this problem are Principal Component Analysis (PCA) [81], Linear Discriminant Analysis (LDA) [82], Nonnegative Matrix Factorisation (NMF), mutual information [83,84], the Pearson correlation coefficient [84,85], the Fisher Discriminant Ratio (FDR) [86], the absolute value of two-sample t-test statistic [62, 65, 74, 87], or the simpler Region Of Interest (ROI) parcellation [77,88]. There are also classifiers with a built-in capacity of feature selection such as LASSO regression [89,90], elastic net [91] and Adabost [92]. One innovative selection procedure has been recently proposed by Bicacro, ETDS, based on the tracking of an expert physician gaze [85]. Morgado extended this approach with TI-ETDS and TD-ETDS, while also testing a mutual information based approach, known as Minimal Redundancy Maximal Relevance (mRMR) [63].

A recent study, driven by Chu et al. [93], brought into light interesting results about the usefulness of feature selection with MRI neuroimaging data using SVM as classifier. The authors concluded that feature selection does improve accuracy when it is based on a priori known ROI’s, regardless of sample size. When based on data-driven approaches, as when using voxels, there is no improvement with feature selection. In both cases there is an improvement when sample size increases. Cuingnet et al. reported similar conclusions in an extensive comparative study based on 10 different approaches [94], underlining the time consumption drawback that comes along with feature selection.

In terms of imaging extracted features, the most used in literature are tissue densities (e.g. gray matter), cortical thickness, shape and volume of hippocampus in the case of MRI, or the uptake of glucose by the brain in PET, both as voxels or average over regions of interest obtained from anatomical atlas [94]. Also, patches of voxels are sometimes used [50], [65]. These are well suited to MIL in the context of AD, since they catch the ambiguity of the disease: not all of them express disease characteristics. Besides, as noticed by Tong et al. [65], given that patches usually share some information, they unveil the inherent structure of images and help diagnosis. Other approaches include texture features, such as 3D Haar-like features, histograms of gradient magnitude and orientation (HGMO) [95] or local binary patterns (LBP) [84].

2.4 CAD studies in AD

There is very little work done on CAD using multiple-instance learning. Concretely, one paper has been written about mammography [67], one about pulmonary embolism [96] and another one in the context of liver cancer [97]. A short review of these papers is presented here.

Briefly, Krishnapuram et al. [67] applied logistic regression under the MIL paradigm to breast cancer diagnosis using the noisy-or model to define the probability that a bag is correctly classified. Bi et al. [96] addressed pulmonary embolism with an innovative approach on computed tomography-angiography images. First, a tobogganing segmentation procedure was applied, in which candidate
pulmonary regions were defined in spatial clusters. Then, a similarity matrix was built for each patient, where each entry represented the distance between the patient’s candidate clusters. This matrix was then used to represent the patient in a 1-norm SVM adapted to the MIL framework. This method, which the authors called Spatial Multiple-Instance Learning, presented 91% sensitivity. In [97], Jiang et al. used abdominal liver CT images and proposed SVM-IOMIL in order to diagnose liver cancer. This instance-level classifier had as input different ROI’s from each image. Each instance was built based on four texture features (angular second moment, entropy, contrast and correlation) extracted from gray level concurrence matrices. Each bag would be classified as positive if the ratio of positive instances was bigger than a given threshold. Additionally, the algorithm was compared to C-KNN and linear SVM, outpacing these two, with 98% accuracy versus 79.3% and 83.2%, respectively.

Regarding CAD applied to AD using MIL, only two papers have been published, written both by the same research team [65, 98]. In [65], Tong et al. developed the first MIL algorithm designed to detect AD and MCI using MRI images from the ADNI database, achieving 88.8% of accuracy between AD patients and healthy controls and 69.6% of accuracy between stable MCI and progressive MCI in a leave-one-out cross-validation procedure. In this paper, each image is segmented so that the hippocampus is used as ROI. Each image is a bag, the instances of which are patches inside the hippocampus. The patch selection method is accomplished by calculating the average p-value of voxels and then selecting the K patches with the lowest mean p-value and satisfying a distance threshold that guarantees less redundancy. SVM was used as classifier, with a graph bag-level kernel called mi-Graph, inspired by work developed by [99] and [100]. This introduced a widespread notion that instances are rarely independent and identically distributed. The graph was implicitly defined by an affinity matrix which took into account the dependency relationship between different patches.

In a following paper, Tong et al. [98] extended their approach to the whole brain and changed their patch extraction method, using a previously implemented penalised regression model together with a resampling scheme [91] which results in probabilities being attributed to features, known as elastic net [101]. The authors obtained 89% accuracy between AD and controls, 82.9% accuracy between controls and progressive MCI and 70.4% accuracy between progressive and stable MCI.

In contrast to MIL, there is a lot of work produced in the field of CAD applied to AD in the last three decades. Stoeckel et al. [102] produced in 2001 a hallmark paper classifying individual subjects. Before that, the majority of studies had performed group analyses and results were focused on average regional abnormalities. It should be noted that some papers, although few, were published before, almost all using neural networks. To the best of our knowledge, the first was published in 1990, by Kippenhan et al. [103], obtaining a remarkable 91% sensitivity and 93% specificity with a data set constituted by 30 controls and 22 AD patients, using PET images. Features inserted into the neural network corresponded to regional glucose consumption from 8 brain regions with the mean removed so that the overall mean was null. In a second paper, Kippenhan et al. [104] further studied Neural Networks with PET, implementing a CAD system to evaluate accuracy in probable AD and possible AD, both versus controls. The areas under ROC curve were 0.85 and 0.81, respectively. In addition, it was assessed whether an expert observation was better. The results were very similar, although slightly
better for the expert.

In [102], Stoeckel et al. used SPECT data containing cerebral blood flow measures from 29 patients with AD and 50 controls. An important aspect is their avant-garde analysis of the curse of dimensionality (referred to as small sample size problem) in AD 3D images when recurring directly to voxels as features instead of some sort of feature extraction, such as symmetry or texture features. In addition, they tested two linear classifiers, the nearest mean classifier (NMC) and the Pseudo Fisher Linear Discriminant (PFLD), the latter providing better results - 89.9% accuracy against 84.8%. On the other hand, they obtained a 82.8% sensitivity against 66.4% for highly experimented observers, a worthy result.

Following this paper, Fung and Stoeckel [105] applied SVM for classification of SPECT images, performing feature selection simultaneously through a 1-Norm SVM version which allows for selection of informative brain areas instead of voxels, thus incorporating spatial information, with 84.4% sensitivity and 90.9% specificity.

In the meantime, Herholz et al. [106] carried out a large-scale study in eight research centres related to AD. FDG-PET images were used and analysed through Statistical Parametric Mapping (SPM). In order to classify the patients, a new quantitative measure was defined, the AD t sum - the sum of t-values in a mask of voxels highly correlated to disease - which was proved to be related to dementia severity and was higher in demented people than in controls. The sensitivity and specificity were both 93%. Besides, a net correlation between the AD t sum and the MMSE score was shown.

After 2008, there was a boost in research in CAD of AD [84]. Among the vast work done since then, Duchesne et al. [83] performed a cross-sectional study based on MRI images to distinguish AD from healthy controls using voxel intensities and deformation fields that included local shape information from a volume of interest smaller than the whole brain. They performed also a feature selection step with PCA. The authors tested three types of classifiers, namely, SVM, LDA and Quadratic Discriminant Analysis (QDA). The best results were obtained with SVM, holding 92% accuracy.

Silveira et al. [92] applied a boosting classifier known as Adabost to differentiate AD, MCI and CN individuals. Adabost belongs to ensemble methods because it combines different classifiers results. This algorithm selects the best features simultaneously with the training step, based on weak classifiers associated each to one feature in an iterative process. Thus, Adabost combines each feature classifier in order to obtain the best weighted accuracy. The results were 90.97% for CN vs. AD, 79.63% for CN vs. MCI and 70% for MCI vs. AD.

In 2010, Cuingnet et al. [94] performed an extensive analysis of Alzheimer’s disease classification based on structural MRI images from ADNI, comparing the performance of ten approaches, from which 5 voxel-wise based methods, 3 cortical thickness based methods and 2 based on the shape and volume of the hippocampus. All the methods were tested in AD vs. CN, CN vs. conversor MCI and conversor MCI vs non conversor MCI. This study aimed at analysing different classification procedures based on the same data, a fact that often limits inter-study comparisons. In addition, the effects of image registration were studied. The authors concluded that using the whole brain is particularly advantageous in CN vs. AD cases, specially the most advanced ones, instead of resorting to the
On the other hand, for intermediate cases, Cuignet et al. suggested the use of a set of selected regions or the hippocampus. Briefly, the results were better for AD vs CN, up to 81% sensitivity and 95% specificity. In terms of CN vs MCIc, the best results were 68% sensitivity and 95% specificity, while for MCIc vs MCInc these are not significantly higher than chance.

Liu et al. [87] tested an hierarchical classification approach, in which different classifiers were disposed in layers. In short, they used Gray Matter (GM) of MRI brain images from ADNI, from which cubic patches were extracted and used as low-level features fed into two low-level classifiers. The outputs of these classifiers were then introduced into high-level classifiers together with statistical measures from patches and the outputs assembled in a final classifier. All the classifiers were chosen to be linear SVM, and the overall accuracy between AD and CN was 91%, while between MCI and CN it was 85.3%.

An innovative procedure was implemented by Coupé et al. in 2012 [82], in which the authors applied the so called SNIPE (scoring by nonlocal image patch estimator), a method in which a subject receives a voxel-wise score based on the similarity of its patch to patches in training subjects from control and demented groups. The average score is then used as the input feature for classification together with ROI volume features. This was done for the hippocampus and entorhinal cortex in an ADNI data set. SNIPE has the advantage of allowing simultaneous segmentation and scoring, and contrasting to atlas-based methods using nonlinear registration, this approach allows inter-subject variability and one-to-many mapping, that is, each subject has its own segmentation. LDA was used and showed that grading biomarkers are substantially better than volumetric ones, with 91% accuracy for CN vs AD, 88% accuracy for CN vs pMCI and 73% for pMCI vs sMCI. Furthermore, the hippocampus alone obtained better results than the entorhinal cortex.

All the papers presented until now covered cross-sectional analyses, that is, were based on a single instant of time. Lately, with increasingly accessible longitudinal data sets, some interest has been given to multi-modal and longitudinal analyses. Chen et al. [75] performed a study where baseline and 12 months Cerebral Metabolic Rate (CMR) of glucose normalised to a spared region of interest were followed throughout that span of time for probable AD patients, MCI patients and healthy controls. The results indicated a significant decline in CMR values for the AD sample, followed by a relevant decrease in MCI and a small decrease in controls. Hinrinchs et al. [107] implemented what they called multi-modal disease marker (MMDM) to study the progression of MCI to AD. Briefly, this method consisted of a combination of multiple imaging modalities and clinical data for classification to input into a SVM classifier using the multi-kernel learning (MKL) framework. In particular, MRI and FDG-PET data together with non-imaging data such as APOE genotype, assays for some CSF compounds and cognitive markers were used. Also, longitudinal data was assessed both through voxel-wise temporal difference and ratio, but the accuracy was not higher than 65% which suggested that dementia effects accumulate slowly enough not to be detected with high accuracy. Accuracy combining all modalities in CN vs. AD was superior to each modality alone and stood at 92.4%.

In an interesting work in 2012, Li et al. [88] presented a longitudinal analysis of cortical thickness changes due to AD in a MRI ADNI data set. In short, the authors first selected the most discrimi-
native cortical ROIs, from which static (baseline and endline cortical thickness measures), dynamical (thinning ratio and thinning speed) and network (clustering coefficient of brain network) features were extracted. Then, minimum redundancy maximum relevance (mRMR) was used to select the best feature subset. Most importantly, results improved considerably when using dynamic and network data, from 89.6% to 96.1% and 76.7% to 80.3%, in CN vs AD and sMCI vs pMCI, respectively. Despite improving the results, it should be noted that dynamic features per se did not have high correlation with subjects labels, which goes in line with work by Hinrichs [107], Rodrigues [108] and Gray [77]. This suggests that dynamic measures give complementary information in AD diagnosis, particularly between the two MCI groups.

More recently, Gray et al. [77] used multi-region FDG-PET information for classification using Yakushev Normalisation and longitudinal information. The authors confirmed that longitudinal data per se is not powerful in the diagnosis, but when concatenated with cross-sectional data it shows to improve accuracy. With only 83 regions they obtained interesting results, more precisely 88% accuracy between AD and controls and 63% between sMCI and pMCI. In a similar but voxelwise approach, Rodrigues [108] obtained also promising results using Yakushev Normalisation and longitudinal information.

Tables (2.1), (2.2) and (2.3) summarise the discussion above, providing information about studies based on MIL, longitudinal studies and supervised approaches, respectively.

### Table 2.1: Performance of different Multiple-Instance Learning CAD systems by chronological order. Results are presented according to the following notation: 1 - CN vs. AD; 2 - CN vs. MCI; 3 - MCI vs. AD; 4 - CN vs. pMCI; 5 - sMCI vs. pMCI; 6 - pMCI vs. AD

<table>
<thead>
<tr>
<th>Authors</th>
<th>Biomarker</th>
<th>Learning Algorithm</th>
<th>Feature Selection</th>
<th>Data Set</th>
<th>acc(%)</th>
<th>sen(%)</th>
<th>spe(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tong et al., 2013 [65]</td>
<td>MRI</td>
<td>mi-Graph</td>
<td>p value reciprocal</td>
<td>231 CN</td>
<td>88.8(1)</td>
<td>85.9(1)</td>
<td>91.3(1)</td>
</tr>
<tr>
<td>Tong et al., 2014 [98]</td>
<td>MRI</td>
<td>mi-Graph</td>
<td>Elastic Net</td>
<td>198 AD</td>
<td>89.0(1)</td>
<td>84.9(1)</td>
<td>92.6(1)</td>
</tr>
</tbody>
</table>

### Table 2.2: Performance of different longitudinal CAD systems by chronological order. Results are presented according to the following notation: 1 - CN vs. AD; 2 - CN vs. MCI; 3 - MCI vs. AD; 4 - CN vs. pMCI; 5 - sMCI vs. pMCI; 6 - pMCI vs. AD

<table>
<thead>
<tr>
<th>Authors</th>
<th>Biomarker</th>
<th>Learning Algorithm</th>
<th>Feature Selection</th>
<th>Data Set</th>
<th>acc(%)</th>
<th>sen(%)</th>
<th>spe(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gray et al., 2012 [77]</td>
<td>PET</td>
<td>SVM</td>
<td>ROI</td>
<td>54 CN</td>
<td>88.4(1)</td>
<td>83.2(1)</td>
<td>93.6(1)</td>
</tr>
<tr>
<td>Rodrigues et al., 2014 [108]</td>
<td>PET</td>
<td>SVM</td>
<td>t statistic</td>
<td>48 AD</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 2.3: Performance of different CAD systems by chronological order. Results are presented according to the following notation: 1 - CN vs. AD; 2 - CN vs. MCI; 3 - MCI vs. AD; 4 - CN vs. pMCI; 5 - sMCI vs. pMCI; 6 - pMCI vs. AD

<table>
<thead>
<tr>
<th>Authors</th>
<th>Biomarker</th>
<th>Learning Algorithm</th>
<th>Feature Selection</th>
<th>Data Set</th>
<th>acc(%)</th>
<th>sen(%)</th>
<th>spe(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kippenhan et al., 1990 [103]</td>
<td>PET</td>
<td>neural networks</td>
<td>ROI</td>
<td>30 CN</td>
<td>91.9(1)</td>
<td>91.0(1)</td>
<td>93.0(1)</td>
</tr>
<tr>
<td>Kippenhan et al., 1992 [104]</td>
<td>PET</td>
<td>neural networks</td>
<td>ROI</td>
<td>100 CN</td>
<td>AUC = 0.85(1)</td>
<td>AUC = 0.81(4)</td>
<td></td>
</tr>
<tr>
<td>Stoeckel et al., 2001 [102]</td>
<td>SPECT</td>
<td>PFLD</td>
<td>Subsampling</td>
<td>50 CN</td>
<td>89.9(1)</td>
<td>82.8(1)</td>
<td>94.0(1)</td>
</tr>
<tr>
<td>Stoeckel et al., 2005 [105]</td>
<td>SPECT</td>
<td>CSVM</td>
<td>CSVM</td>
<td>31 CN</td>
<td>86.0(1)</td>
<td>84.1(1)</td>
<td>90.9(1)</td>
</tr>
<tr>
<td>Duchesne et al, 2008 [83]</td>
<td>MRI</td>
<td>SVM</td>
<td>-</td>
<td>75 CN</td>
<td>92.0(1)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Silveira et al, 2010 [92]</td>
<td>PET</td>
<td>Adabost</td>
<td>Adabost</td>
<td>81 CN</td>
<td>91.0(1)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bicacro, 2011 [85]</td>
<td>PET</td>
<td>SVM</td>
<td>ETDS</td>
<td>59 CN</td>
<td>91.4(1)</td>
<td>90.0(1)</td>
<td>92.8(1)</td>
</tr>
<tr>
<td>Morgado et al., 2013 [63]</td>
<td>PET</td>
<td>SVM</td>
<td>mRMR</td>
<td>59 CN</td>
<td>87.9(1)</td>
<td>84.9(1)</td>
<td>90.8(1)</td>
</tr>
<tr>
<td>Liu et al., 2014 [87]</td>
<td>MRI</td>
<td>Hierarchical Fusion</td>
<td>t statistic</td>
<td>229 CN</td>
<td>92.0(1)</td>
<td>90.9(1)</td>
<td>93.0(1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>225 MCI</td>
<td>85.3(2)</td>
<td>82.3(2)</td>
<td>88.2(2)</td>
</tr>
</tbody>
</table>
3

Multi-Instance Learning - A theoretical overview

Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>24</td>
</tr>
<tr>
<td>3.2 State of the art</td>
<td>24</td>
</tr>
<tr>
<td>3.3 Mathematical Background</td>
<td>28</td>
</tr>
</tbody>
</table>
3.1 Introduction

MIL, as supervised learning, involves labeling unseen examples after a training process with labeled examples through the inference of a function that takes into account the relationship between these and their label. However, MIL differs from the latter in that each example may have more than one feature vector contained inside. Great attention due to its theoretical interest and broad applicability has been given in the past few years.

MIL is applicable both in classification and regression problems. Notwithstanding, the focus in this chapter will be only on the first task, and in particular the binary case, since this thesis is based on this paradigm.

This chapter gives the reader a theoretical background over MIL. The reader is first presented with an outline of the state of the art, becoming acquainted with some of the work done on the area and general concepts. Then, in section (3.3), the mathematics behind is explained in detail.

3.2 State of the art

The term Multiple-Instance Learning (also termed multi-instance learning) first appears in 1997, in the foremost paper written by Dietterich et al., being defined as tasks where the training examples are ambiguous: a single example object may have alternative feature vectors (instances) that describe it, and yet only one of those feature vectors may be responsible for the observed classification of the object [109].

Conceptually, MIL lies in the middle between supervised learning and unsupervised learning, since there is an intrinsic ambiguity in instances, which are grouped in multisets (usually called bags). While instances labels are unknown, those of bags are known a priori. Some authors consider MIL to be a special case of semi-supervised learning [110], or even of supervised learning, the only nuance being the nature of training examples [111,112].

Figure 3.1: Supervised Learning and Multiple Instance Learning. Source: [109].
The founding work led by Dietterich et al. \cite{109} was motivated by a drug activity prediction study concerning musk molecules, which in some conformations are detected by human nose with smell. It was assumed each molecule was a bag and each of its many conformations was an instance. Molecular conformations were measured in a ray-based representation, where each ray was emanated from the molecule centre. These, plus four features associated to the position of an oxygen atom, made up to 166 features per instance. The positive bags (termed musk molecules) had at least one positive conformation (instance), the negative ones (termed non-musk molecules) should have no positive instances. This corresponds to an asymmetric and somehow restrictive assumption commonly called the Standard Multi-Instance (SMI) assumption. The asymmetry lies on the fact that if we change the criteria to classify positive and negative bags, the labels will have a different interpretation \cite{114}.

Dietterich et al. created the Axis Parallel Rectangle (APR) algorithm. This relies on the shrinking or expansion of an hyper-rectangle so that the number of positive points (instances) inside is maximised, while minimising the number of negative points inside. Any bag that contains an instance within the rectangle is classified as positive. This algorithm achieved the best accuracy up to date with the Musk1 and Musk2 data sets from the UCI machine learning repository \footnote{Available at \url{http://archive.ics.uci.edu/ml/}} (91.4 % and 89.2 % accuracy, respectively), but has been criticised for being specially designed for it, thus not generalizing well to other problems. An important aspect, proven by Long and Tan, is that APR is Probably Approximately Correct (PAC) learnable \cite{115}. However, in this derivation instances were considered to be independent and the time bound was very high. Auer et al. then showed that if instances are dependent, the problem becomes NP hard \cite{116}, and Blum and Kalai proved that the MIL framework is PAC-learnable with one-sided random classification noise \cite{117}.

Following the SMI assumption, Maron \cite{113} developed the next year a leading algorithm, DD, built on a probabilistic approach that searched for the concept in the instance space that distinguished the most between positive and negative instances. Its accuracy in the Musk datasets was 89.5% and 82.5%, respectively. This work would be the seed for following works such as EM-DD \cite{118} and DD-SVM \cite{119}. Briefly, EM-DD uses the expectation maximisation algorithm and DD. Contrarily to DD, which searches for the maximum concept in the space defined by all "positive" instances (all instances belonging to positive bags), EM-DD turns the problem into a single-instance problem by choosing only one positive instance per positive bag, which introduces computational efficiency advantage when compared to DD. Although its reported results were better than DD in the Musk datasets, some authors argued that these had been obtained based on the best hypothesis in the test set. The
correction showed the algorithm’s accuracy was not superior [114,120]. On the other hand, DD-SVM uses DD to define instance prototypes and then applies a nonlinear mapping defined based on these prototypes. Finally, a SVM classifier is used in the new space. This algorithm may be considered a direct antecessor of MILES, and consequently, also of YARDS [121].

Not long after Maron formulation, Andrews et al. [112] presented two maximum margin solutions inserted in the MIL paradigm, the maximum pattern margin formulation of MIL and maximum bag margin formulation of MIL, which correspond to two algorithms, mi-SVM, and MI-SVM, respectively. Both algorithms are mixed integer programming problems in which there is an optimal hyperplane and an optimal labeling to be found, for which the authors presented an optimization heuristic. The results achieved on the Musk data set were competitive. Also resorting to SVM, Zhou and Xu [110] proposed a semi-supervised approach to solve the MIL problem, named missSVM (Multi-Instance learning by Semi-Supervised Support) (see table (3.1)). Worthy of note, in a different approach, Ramond and Raedt [122] extended neural networks with backpropagation to multi-instance learning through the introduction of subnetworks that tackle the instance-level classification and smoothed max nodes, achieving remarkable results [122].

In 2003, Weidmann [123] introduced three innovative assumptions concerning the relationship between bag’s label and its instances’ labels: the presence-based assumption, the threshold-based assumption and the count-based assumption, which are in order of ascending generalisation. In the SMI assumption there is a single concept in the instance space and the requirement is that at least one instance belongs to it to be positive. In contrast, this generalisations allow for the existence of several concepts, which can be interpreted as the existence of several classes of instances. The first one states that there should be at least one positive instance for each concept inside a bag. This means that the existence of positive instances is necessary but not sufficient per se. The SMI assumption is a particular case, since there is only one concept. The second assumption extends the first by simply requiring a threshold-value for the number of positive instances per concept in a bag. Finally, the last assumption introduces an upper threshold to the number of positive instances.

In an interesting review, Ray and Craven [124] compared MIL with supervised learning. This study led to some important conclusions, namely that no MIL algorithm is superior to its supervised counterpart in every domain, and in some cases the supervised algorithms are superior. The authors also proposed a logistic regression based MIL algorithm, MI-LR, which attained competitive results in comparison with literature. Also on the relationship between MIL and supervised learning, Zhou and Zhang showed the latter can be adapted into the multiple instance framework, provided that the focus on discrimination lies at the bag level [125]. This is indeed the case of CKNN, a nearest neighbour algorithm based on the well-known k-NN. The shift to bags is accomplished through a new definition of distance between these entities, the minimal Hausdorff distance. For some time, CKNN held the best results in the Musk1 data set [126].

In a recent and key paper, Amores [127] formulated a comparative study in which he considered the existence of three paradigms in MIL classification:

- Instance-Space paradigm - The discriminative information lies on instances, that is, the classifier
receives instances as input. Some of the algorithms developed under this paradigm are DD [113], Expectation-Maximization Diverse Density [128], and both MI-SVM and mi-SVM [112]. The first MIL algorithms belonged to this paradigm.

- Bag-Space paradigm - The classifier receives bags as input. It generally relies on the definition of a distance between bags. Some common distances are the minimal Hausdorff distance, Chamfer distance or Earth-Movers distance. CKNN [126] represents fairly well this model. MI Kernel, a bag kernel that may be used with Support Vector Machines, is also a good example, developed by Gartner [129].

- Embedded-Space paradigm - As in the previous paradigm, there is information extracted directly from bags but in an explicit way, by mapping the bag from the bag space to a single instance vector space. One example is the algorithm simpleMI, also called statistical kernel, in which a bag is represented by statistical measures such as maxima or minima over features in a single vector [120]. MILES, YARDS and BARTMIP are also paradigmatic of this case.

This nomenclature is complete, in the sense that every published MIL algorithm falls into one of the paradigms. All these topics are discussed in deeper detail in section (3.2), where the mathematical theory behind MIL will be presented. Besides this comprehensive study, Amores also developed a new embedded space based algorithm, named MILDE (Multiple Instance Learning by Discriminative Embedding), which joins aspects both from this paradigm and the instance space through the existence of instance-level classifiers whose outputs are then fed into bag-level classifiers and then mapped to a new space [130].

MIL has ever since been used in many research areas, such as classification of molecules in the context of drug design, content-based image retrieval, face recognition and text categorisation [112]. Besides the two popular musk data sets, some other databases have been used, such as the COREL photo library (containing three data sets: elephant, fox and tiger), tested by Maron and Ratan [131], Andrews et al. [112] and Zhou et al. [99], among other authors, or the TREC9 text data set, tested by Andrews et al. [112].

Table (3.1) presents accuracy results for some of the benchmark MIL algorithms, and also for a supervised one, tested by Ray et al. [124], by simply attributing to instances their bag label. Noteworthy, Zhou and Zhang [125] analysed the effect of using ensembling within MIL framework, and obtained significantly better results with the Musk data sets in all algorithms tested: APR, DD, CKNN and EM-DD.
Table 3.1: Accuracy results on some of the benchmark MIL algorithms in the two Musk and three COREL data sets. * - values presented refer to area under the ROC curve.

<table>
<thead>
<tr>
<th>Classification Algorithm</th>
<th>Musk1</th>
<th>Musk2</th>
<th>Elephant</th>
<th>Fox</th>
<th>Tiger</th>
<th>Evaluation</th>
<th>Paradigm</th>
</tr>
</thead>
<tbody>
<tr>
<td>APR [109]</td>
<td>91.4%</td>
<td>89.2%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>10-fold CV</td>
<td>Instance Space</td>
</tr>
<tr>
<td>DD [113]</td>
<td>89.5%</td>
<td>82.5%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>10-fold CV</td>
<td>Instance Space</td>
</tr>
<tr>
<td>DD (ensembled) [125]</td>
<td>91.8%</td>
<td>89%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>10-fold CV</td>
<td>Instance Space</td>
</tr>
<tr>
<td>MULTINST [116]</td>
<td>76.7%</td>
<td>84.0%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>10-fold CV</td>
<td>Instance Space</td>
</tr>
<tr>
<td>EM-DD [128]</td>
<td>84.8%</td>
<td>84.9%</td>
<td>75.9%</td>
<td>61.3%</td>
<td>N/A</td>
<td>10-fold CV</td>
<td>Instance Space</td>
</tr>
<tr>
<td>EM-DD (ensembled) [125]</td>
<td>96.9%</td>
<td>97%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>10-fold CV</td>
<td>Instance Space</td>
</tr>
<tr>
<td>DD-SVM [119]</td>
<td>85.8%</td>
<td>91.3%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>10-fold CV</td>
<td>Embedded Space</td>
</tr>
<tr>
<td>mi-SVM [112]</td>
<td>87.4%</td>
<td>83.6%</td>
<td>82.2%</td>
<td>58.2%</td>
<td>78.9%</td>
<td>10-fold CV</td>
<td>Instance Space</td>
</tr>
<tr>
<td>MI-SVM [112]</td>
<td>77.9%</td>
<td>84.3%</td>
<td>81.4%</td>
<td>59.4%</td>
<td>84.0%</td>
<td>10-fold CV</td>
<td>Instance Space</td>
</tr>
<tr>
<td>Citation-KNN [126]</td>
<td>92.4%</td>
<td>86.3%</td>
<td>50.0%</td>
<td>50.0%</td>
<td>50.0%</td>
<td>10-fold CV</td>
<td>Bag Space</td>
</tr>
<tr>
<td>Citation-KNN (ensembled) [125]</td>
<td>94.8%</td>
<td>87.1%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>10-fold CV</td>
<td>Bag Space</td>
</tr>
<tr>
<td>Bayesian-KNN [126]</td>
<td>90.2%</td>
<td>82.4%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>10-fold CV</td>
<td>Bag Space</td>
</tr>
<tr>
<td>simpleMI</td>
<td>89.6%</td>
<td>84.7%</td>
<td>77.3%</td>
<td>60.3%</td>
<td>73.6%</td>
<td>10-fold CV</td>
<td>Embedded Space</td>
</tr>
<tr>
<td>MI Kernel [129]</td>
<td>86.4%</td>
<td>88.0%</td>
<td>84.3%</td>
<td>60.3%</td>
<td>84.2%</td>
<td>10-fold CV</td>
<td>Bag Space</td>
</tr>
<tr>
<td>BARTMIP [132]</td>
<td>94.1%</td>
<td>89.8%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>LOO CV</td>
<td>Embedded Space</td>
</tr>
<tr>
<td>MILES [133]</td>
<td>86.3%</td>
<td>87.7%</td>
<td>84.3%</td>
<td>61.0%</td>
<td>82.0%</td>
<td>10-fold CV</td>
<td>Embedded Space</td>
</tr>
<tr>
<td>miGraph [99]</td>
<td>90.0%</td>
<td>90.0%</td>
<td>86.8%</td>
<td>61.6%</td>
<td>86.0%</td>
<td>10-fold CV</td>
<td>Bag Space</td>
</tr>
<tr>
<td>MIIGraph [99]</td>
<td>88.9%</td>
<td>90.3%</td>
<td>85.1%</td>
<td>61.2%</td>
<td>81.9%</td>
<td>10-fold CV</td>
<td>Bag Space</td>
</tr>
<tr>
<td>YARDS [114]</td>
<td>84.4%</td>
<td>82.6%</td>
<td>83.9%</td>
<td>55.9%</td>
<td>81.3%</td>
<td>10-fold CV</td>
<td>Embedded Space</td>
</tr>
<tr>
<td>MissSVM [110]</td>
<td>87.6%</td>
<td>80.0%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>10-fold CV</td>
<td>Instance Space</td>
</tr>
<tr>
<td>MLDE [130]</td>
<td>87.1%</td>
<td>91.0%</td>
<td>85.0%</td>
<td>66.5%</td>
<td>83.0%</td>
<td>10-fold CV</td>
<td>Embedded Space</td>
</tr>
<tr>
<td>mi-NN [122]</td>
<td>88.0%</td>
<td>82.0%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Instance Space</td>
</tr>
<tr>
<td>MI-LR [124]</td>
<td>0.87</td>
<td>0.87</td>
<td>0.93</td>
<td>0.63</td>
<td>0.93</td>
<td>10-fold CV</td>
<td>Instance Space</td>
</tr>
<tr>
<td>SVM (quadratic) [124]*</td>
<td>0.90</td>
<td>0.85</td>
<td>0.83</td>
<td>0.58</td>
<td>0.83</td>
<td>10-Fold CV</td>
<td>Supervised</td>
</tr>
</tbody>
</table>

Although multiple instance regression, as opposite to multiple instance classification, is out of the scope of this thesis, it should be noted that some research was done on this subject. In this alternative scenario, the label associated to each bag is a real value instead of a discrete one [127]. The first paper referring to MIR was presented by Ray and Page in 2001 [134], revealing the interest of looking into drug activity as a measurable real number. Their approach was tested only in synthetic data. Following their work, little research has been done on the area.

### 3.3 Mathematical Background

In the founding paper by Dietterich et al. [109] there was the objective of predicting specific characteristics of drug molecules. Each example molecule was considered to be a bag of feature vectors and was given a label. The labels of its possible conformations (instances) were unknown. This led to the definition of a new abstraction, bags.

A bag may be defined mathematically as a multiset, a generalisation of the set concept. Whereas a set is defined as a collection of elements with no repetition, in a multiset elements can appear more than once and in any order. In the vast majority of literature, however, bags are treated as sets, which is a simpler assumption but with few practical effects, since instances are vectors of real components and repetitions will happen very rarely [130]. Following this assumption, a bag $B$ is defined as:

$$B = \{\bar{x}_1, ..., \bar{x}_N\}$$ (3.1)
where $\vec{x}_j \in \mathbb{R}^d$ are d-dimensional instances (also termed feature vectors) that live in a space $\chi$, known as the instance space, and $N = |B|$ is the cardinality of $B$, which may vary from bag to bag. As the reader can easily prove, when $N = 1$, the MIL problem becomes the usual supervised one.

Now, let $\Omega = \{0, 1\}$ be a binary class attribute. In an analogy with the supervised learning paradigm [111], we define an instance level concept $\nu_I$ as follows:

$$\nu_I : \chi \rightarrow \Omega$$

(3.2)

There are $2^{|\chi|}$ concepts on $\chi$. If bags are considered to be subsets of $\chi$, it is possible now to define a set concept $\nu_{\text{set}}$ as:

$$\nu_{\text{set}} : 2^\chi \rightarrow \Omega$$

(3.3)

where $2^\chi$ represents the set of all functions from $\chi$ to $\{0, 1\}$, which is the same as to say the set of indicator functions over $\chi$. It should be noted that this formalisation corresponds to bags viewed as sets, although a more general one does exist, namely that proposed by Foulds [111], where $\nu_{\text{set}} : \mathbb{N}^\chi \rightarrow \Omega$, accounting for the multiset nature of multi-instance examples.

Such concepts are sometimes named multi-part concepts [100]. It is easy to see that there are $2^{2^{|\chi|}}$ different multi-part concepts on sets. On the other hand, as defined by Gartner [100], MI concepts are a special kind of these, defined as:

$$\nu_{\text{MI}}(B) \iff \exists \vec{x} \in B : \nu_I(\vec{x})$$

(3.4)

where $\nu_I$ is a concept over an instance space and $B \subseteq \chi$ is a set. In this case there are $2^{|\chi|}$ different concepts. The task of MIL is to learn this function based on a number of examples of this function. A function $f : \chi \rightarrow \mathbb{R}$ is said to separate the instance-level concept if $f(\vec{x}) > 0 \iff \nu_I(\vec{x})$. Similarly, a function $F : 2^\chi \rightarrow \mathbb{R}$ separates the bag-level concept if $F(B) > 0 \iff \nu_{\text{MI}}(B)$. It is this function that, in fact, will be learnt by MIL algorithms. For ease of notation, it will be henceforth be notated as $F$. Also, instance-level classifiers will be noted as $f$.

In order to estimate $F$, a training set of dimension $M$, $T = \{(B_1, y_1), \ldots, (B_M, y_M)\}$, where $y_i \in \{0, 1\}$ is the label of bag $i$ ($y_i = 0$ if the bag’s label is negative, $y_i = 1$ otherwise) is used.

Figure 3.3: An illustration of bags and instances. Source: [130].
The way F is computed is a central topic in MIL, which Foulds [111] and Amores [127] thoroughly analysed, the first focusing mainly on the relationship between bag labels and instances and the latter emphasising how discriminant information is extracted - that is, from bags or from instances - and how bags are represented. These two works constitute up to the moment the only two systematising the nomenclature of MIL and do not conflict with each other, only the perspective changes. In the following lines, Foulds' approach is presented, and then, Amores framework is looked into. It should be noted that many authors in the literature do not make clear which paradigm they follow.

### 3.3.1 Foulds’ Taxonomy

Foulds developed its work essentially based on two previous works elaborated by Xu et al. [135] and Weidmann et al. [123]. The first one divided MIL into two large domains according to the way information is extracted from data: the Instance-based domain and the Metadata-based domain. In the first domain, instances are first classified through an instance-level classifier and only then bags are classified. On the other hand, Metadata approaches assume that some meta-level information describes the examples and map the bags to a single-instance feature space, where the features have some metadata extracted from bags.

Inside the first approach, Foulds inserted Weidmann’s framework, which established the existence of an hierarchy in instance-level classification. This framework generalised the simpler Standard MI assumption, where a bag could be classified simply by the disjunction of instance-level class labels:

$$F(B) \iff (f(\vec{x}_1) \lor f(\vec{x}_2) \lor \ldots \lor f(\vec{x}_N)) \quad (3.5)$$

When $f(\vec{x}) \in \{0, 1\}$ this is equivalent to:

$$F(B) \iff \max_{\vec{x} \in B} f(\vec{x}) \quad (3.6)$$

The reader may understand this with an informal illustration proposed by Chevaliere and Zucker [136], termed the simple jailer problem. Accordingly, let there be a locked door and a set of N bunches of keys. A bunch (bag) is said to be useful, or positive, if at least one of its keys (instances) opens the door, and useless, or negative, otherwise. The learnt model intends to predict whether a keychain is useful or not.

**Figure 3.4:** Learning under the SMI assumption. The classification of each bag, described by N instances, may be seen as an OR logical operator over the learnt concepts of each instance. Adapted from [136].
Weidmann introduced three assumptions, in ascending order of generality: presence-based assumption; threshold-based assumption and count-based approach. The presence-based assumption simply requires a bag to have one or more instances belonging to the existing instance-level concepts. Formally, let \( \nu_{PB} : \mathbb{N}^x \rightarrow \Omega \) be a presence-based MI concept, \( \hat{C} \in C \) the set of required concepts and \( \Delta : 2^x \times C \rightarrow \mathbb{N} \) a function that counts the number of instances belonging to a concept in a bag. Then:

\[
\nu_{PB}(B) \iff \forall c \in \hat{C} : \Delta(B, c) \geq 1 \tag{3.7}
\]

It is easy to see that under the SMI assumption, \( |\hat{C}| = 1 \), that is, only one concept is needed. Using again the jailer metaphor, the presence-based assumption would introduce multiple locks in the door instead of one, and it would be necessary at least one key able to open each lock.

Now, for the threshold-based assumption, let \( \nu_{TB} : 2^x \rightarrow \Omega \) be a threshold-based MI concept. Using the same terminology:

\[
\nu_{TB}(B) \iff \forall c_i \in \hat{C} : \Delta(B, c_i) \geq t_i \tag{3.8}
\]

where \( t_i \in \mathbb{N} \) is a threshold for each concept. This translates into the requirement that each bag has at least \( t_i \) instances for concept \( i \). In the jailer allegory, this corresponds to a situation where the door has multiple locks of each type, and they unlock only when there is an equal number of necessary keys.

Finally, the count-based approach introduces a minimum and a maximum value for the number of instances belonging to a given concept:

\[
\nu_{CB}(B) \iff \forall c_i \in \hat{C} : t_i \leq \Delta(B, c_i) \leq z_i, \quad t_i, z_i \in \mathbb{N} \tag{3.9}
\]

This scenario is very similar to the one presented before, except that there is a limit on the number of keys that may be used to unlock the door.

Figure 3.5 represents the MI assumptions explained above, also known as Generalised MI assumptions [123]:

---

**Figure 3.5:** Weidmann’s hierarchy. Source: [111].
3.3.2 Amores’ Taxonomy

Contrarily to Foulds, Amores did not focus on the relationship between instances labels and bags labels. Instead, he analysed the problem in terms of how the information is extracted from these two entities. While some nomenclature remains similar, there are new terms that appear or are more explicitly presented. Under Amores formulation, three main domains arise: the Instance Space (IS) paradigm, the Bag Space (BS) paradigm and the Embedded Space (ES) paradigm. As it will soon be seen, the first one corresponds to Foulds instance based-domain and the latter to the Metadata one.

Methods following the Instance Space paradigm try to infer an instance-based classifier $f(\vec{x}) \in [0,1]$ from the training data and obtain the bag-level classification $F(B) \in [0,1]$ through an aggregation operator such as the OR one (see equation (3.10) and figure (3.6)). The difficulty here lies on the absence of instance labels. Some assumptions must be made about, which led Amores to split this paradigm into two sub-paradigms: the SMI assumption and the collective assumption.

The first one has already been explained above. Many approaches have been proposed, mainly calling upon Support Vector Machine adaptations, such as mi-SVM and MI-SVM, both developed by Andrews et al. [112], or SMIL, by Bunescu et al. [137]. All of them circumvent the absence of labels at the instance level by initialising instances with their bag’s label. The interested reader should consult these references. Two notable exceptions to the use of SVM under this paradigm are the DD algorithm, which is used in this thesis and analysed in depth in chapter 4, and EM-DD [128].

Amores introduces a normalisation factor in (3.5) and generalises it to any operator:

$$F(B) \iff \frac{(f(\vec{x}_1) \circ f(\vec{x}_2) \circ ... \circ f(\vec{x}_N))}{|Z|}$$

where $\circ$ denotes an aggregation operator and $Z$ is a scaling factor, such as the number of instances per bag.

![Figure 3.6: Illustration of the instance-space paradigm. Two bags are shown, both having two classes. It can easily be seen that what makes a bag positive is the presence of class 1 instances, which makes the SMI assumption suitable (a). A classifier is then applied (b). Source: [127].](image-url)
proposed by Bunescu et al. [137] and simply trains a supervised classifier on a training set where the instances inherit their bag’s label and then the bag level classifier is obtained through the average of instance-level classifiers. In order to give the reader a more tangible example, it can easily be seen, referring to (3.5), that the normalisation factor in the last case is the bag size and the operator $o$ is the sum:

$$F(B) = \frac{1}{|B|} \sum_{\vec{x} \in B} f(\vec{x})$$  \hspace{1cm} (3.11)

In an attempt to further the discrimination, another subparadigm emerged to include the so called Weighted Collective methods. It is nothing more than a generalisation of the collective assumption where each instance has its own weight. A general expression for the bag classifier can be presented as:

$$F(B) = \frac{\sum_{\vec{x} \in B} w(\vec{x}) f(\vec{x})}{\sum_{\vec{x} \in B} w(\vec{x})}$$  \hspace{1cm} (3.12)

Two algorithms close to this definition are MILES [133] and YARDS [111], both implemented on this thesis.

Taking a look now at the BS paradigm, it differs from the preceding one in that the learning process is done in the space of bags. This represents a shift in perspective that consists in regarding bags as a whole entity and treating instances homogeneously [138] (see figure (3.7)).

A usual procedure when learning under this paradigm is to use information such as distance between bags and then insert it in distance-based algorithms such as K-NN. However, it should be noted that bags being non-vectorial entities, this means a new definition of dissimilarity must be given to accommodate this model [139]. Several definitions have been proposed, such as the well known Hausdorff distance, commonly used in pattern recognition, or the Earth-Movers distance. The first was adapted and used by Wang and Zucker [126] with an K-NN based approach known as CKNN, an algorithm applied in this thesis. Nonetheless, not all dissimilarities may be plugged into a kernel, for they may not be Euclidean nor metric [140].

Kernels are, in fact, a viable alternative in the definition of a distance measure between bags. Particularly, in a detailed study by Gartner et al. [129], a class of set kernels is defined whose general expression is given by:

$$k_{set}(X, X') = \sum_{\vec{x} \in X, \vec{x}' \in X'} k_\chi(\vec{x}, \vec{x}')$$  \hspace{1cm} (3.13)

where $X$ and $X'$ are two different sets and $k$ is a usual instance kernel in $\chi$, such as the radial basis function.

A careful inspection of equation 3.13 shows that a normalisation factor is needed. Indeed, sets with a larger cardinality will have a higher sum. Gartner proposed the following general expression:

$$k(X, X') = \frac{k_{set}(X, X')}{f_{norm}(X)f_{norm}(X')}$$  \hspace{1cm} (3.14)
where $f_{norm}$ is a normalisation function which may be chosen from various options. Kernels following this expression are Mercer kernels [129].

Still in this work, Gartner developed two specific kernels, the MI kernel and the statistic kernel, which the interested reader is referred to for more details [129]. In addition, Gartner showed that MI concepts are linearly separable with MI kernel provided that instance-level concepts are also separable. Consequently, MI concepts respecting the SMI assumption are learnable with these kernels [111].

Interestingly, as Amores notes, a distance can be inserted into the definition of a kernel, such as the extended Gaussian kernel:

$$k(X, X') = \exp(-\gamma D(X, X'))$$ (3.15)

Inversely, a distance metric may be defined based on a kernel [129]:

$$D(X, X') = \sqrt{K(X, X) - 2K(X, X') + K(X', X')}$$ (3.16)

![Figure 3.7: Illustration of the bag-space paradigm. Training bags are learnt through the definition of a distance (a). A classifier is then applied (b). Source: [127].](image)

Finally, Amores presented what he termed the ES paradigm. Broadly speaking, the idea behind it is the same as the BS space paradigm. However, while the latter extracted information from bags in an implicit way - using a kernel or a distance measure - this paradigm does the same explicitly through a mapping (also known as embedding) from the bag space to a feature vector space. This vector should contain useful information about the bag. According to the approach taken regarding information extraction, one can divide this paradigm in ES methods without vocabulary and vocabulary ES based methods. The first relies only on statistical data, such as the average of instances inside a bag [120], while the vocabulary-based methods are much richer and prone to multiple strategies. The common denominator to all of them is the presence of a vocabulary storing classes (prototypes) of instances, a mapping function and a standard supervised classifier. The learner then uncovers which prototypes are associated with positive and negative bags, conferring different relevance to each type of instances [138].

Formally speaking, a vocabulary is a set with $k$ concepts, each defined by a set of parameters:

$$V = \{(C_1, \theta_1), \ldots, (C_k, \theta_k)\}$$ (3.17)
where $C_j$ corresponds to the concept $j$ and $\theta_j$ the parameters that describe it. The concepts may be interpreted as classes of instances, generally obtained through some clustering method. That is indeed the case of BARTMIP [132], another one of the algorithms used in this thesis.

After the establishment of a vocabulary, a mapping function is applied:

$$M(B, V) : B \rightarrow \vec{v}, \vec{v} = (v_1, ..., v_k)$$

(3.18)

where $B$ denotes a bag and $V$ a vocabulary. Now the bag is represented by a $k$-dimensional vector in the feature space. This embedding process is generally made taking into account the matching between instances $\vec{x}_i \in B$ and the concepts of the vocabulary. Finally, a supervised classifier is applied in the ES using a training set already mapped to this space. The bag classifier may now be defined as:

$$F(B) = f(M(B, V))$$

(3.19)

where $f(\vec{v}) \in [0, 1]$ is the supervised classifier. Equation (3.19) synthethises the idea of ES paradigm: a bag is first embedded in a vectorial space, based on a given vocabulary, and then classified with some supervised classifier such as support vector machines, neural networks or distance-based ones (see figure (3.8)). The embedded space, when resulting from a distance based mapping, is also known as dissimilarity space [140].

Figure 3.8: Illustration of the embedded-space paradigm. Training bags are mapped into an instance space where they correspond to points as in usual supervised learning (a). A supervised classifier is then applied (b). Source: [127].
4

Methodology

Contents

4.1 Introduction ........................................... 38
4.2 Image preprocessing .................................... 38
4.3 Yakushev Normalisation ................................ 39
4.4 Feature Extraction and Selection ....................... 40
4.5 Classification .......................................... 42
4.6 Longitudinal Analysis .................................. 63
4.7 Simulation of non-registered images .................... 63
4.1 Introduction

This chapter aims at presenting to the reader a detailed description of the methodology carried out through this work. Image preprocessing steps are first explained. Although being also a preprocessing step, Yakushev Normalisation is presented in a separated subsection due to its relevance. The ensuing topic is devoted to feature extraction and selection, and then, all the learning algorithms applied in this thesis are analysed in detail: DD, CKNN, MILES, YARDS, BARTMIP, and also the supervised counterpart, SVM, for comparison purposes.

The main idea behind the use of these algorithms was the implementation of at least one method in each paradigm of MIL (IS, BS and ES). DD is one of the most paradigmatic algorithms and belongs to the first paradigm. CKNN depicts in a very intuitive way the Bag Space paradigm. Finally, the last three explore different approaches inside the ES paradigm, which is considered in the literature as the most promissory approach. MILES and YARDS are very similar in nature and map bags to a new space where each feature is defined by the distance to training instances. BARTMIP implements an unsupervised based classifier, which is on its own right an interesting perspective. Each of these methods was initially tested with the Musk data sets and compared to literature results in order to validate them.

The chapter ends with a brief overview about the longitudinal analysis and simulation of non-registered images classification, both tested recurring to MILES.

4.2 Image preprocessing

ADNI data comes from a wide range of different FDG-PET scanners, which introduces undesired variations in image obtention. In order to obtain more uniform and similar FDG-PET data, it was submitted to four pre-processing steps carried out by ADNI [49]:

1. Dynamic Co-registration - Raw images were converted to a standard file format (DICOM). Either six five-minute frames or four five-minute frames were acquired 30 to 60 minutes post-injection and each was co-registered to the first one. The base frame image and the others were then recombined to a co-registered dynamic image set. These image sets had the same image size and voxel dimensions, while maintaining the spatial orientation of the original PET image. This step decreases the patient motion artifacts.

2. Average - All frames were averaged, resulting in a single PET image per patient.

3. Image and voxel size standardisation - Each averaged dynamic co-registered image was reoriented into a voxel image grid of size 160x160x96, with 1.5 mm$^3$ cubic voxels where the anterior-posterior axis of the subject was parallel to the AC-PC line. Each subject’s image obtained in this way was used as a reference on that subject. Also, using a subject specific mask, the average of voxel intensities inside the mask was set to one, which corresponds to a cerebral global mean normalisation. This allowed a better comparison between different scanner models.
4. Resolution uniformisation - Each image was filtered to produce images of a uniform resolution of 8 mm FWHM, which corresponds to the lowest resolution scanners used in ADNI.

After ADNI preprocessing steps, images from different patients were still not aligned between them, which rendered voxelwise intersubject comparison impossible. This was circumvented via a warping into the MNI152 standard space, as described in Morgado and Silveira [141]. First, brain tissue in all MR images was extracted (skull-stripping), and classified into white matter and gray matter through SPM8, a software capable of producing probability maps for each of these tissues. Then, these images were segmented into white matter and gray matter, recurring to FreeSurfer, an open source software to analyse MRI brain images. Then, all PET images were co-registered with the skull-stripped MR images using SPM8, through rigid-body transformations with 6 degrees of freedom (3 rotations plus 3 translations) and an objective function based on the normalised mutual information between images. All images acquired on different months were next non-linearly registered into a subject-specific template using the DARTEL toolbox from SPM8. Finally, all MR images were non-linearly registered to an intersubject template using the aforementioned toolbox. This template was then mapped to the MNI-ICBM 152 non-linear symmetric atlas (version 2009a), using an affine transformation. After the completion of these steps, original PET images were resampled into the MNI152 standard space with a 1.5x1.5x1.5 mm resolution. All this process is summarised in figure (4.1).

Figure 4.1: Image registration steps. Source: [141].

4.3 Yakushev Normalisation

The VI features used in this thesis were obtained from FDG-PET images retrieved from the ADNI database [22] and represent a direct measure of the glucose uptake detected in a given voxel. ADNI data had already undergone a series of preprocessing steps that co-registered and normalised images (see section (4.2)). This allowed each one to have the same dimension in the MNI space (121x145x121) and to perform voxelwise comparisons. As a result, each patient data contained a number of features in the order of millions, most of them belonging to areas outside the brain.

Before any feature selection, an adaptation of Yakushev normalisation [68] was applied. The significance of this step lies on the extinction of the false hypermetabolised areas artifact, which appear when images are normalised to the cerebral global mean (see chapter 2). Briefly, as performed

---

1Available at http://freesurfer.net/.
2This is the template adopted by the International Consortium for Brain Mapping (ICBM), which is an average of 152 normal MRI scans that were matched to a previous template, the MNI305.
in Rodrigues [108], a SPM like approach was taken. A paired t-test was performed to provide a
voxelwise computation of t statistics. Then, in order to obtain a reference cluster of voxels, this step
required the setting of two parameters - a t-value threshold (t) and a minimum cluster radius (r) (see
table (4.1)). From those voxels satisfying these constraints, only the cluster with the highest mean t-
value was selected. The mean of voxel intensities inside the cluster was then used as the normalisation
factor. The best parameters for each modality (CN vs. MCI, CN vs. AD and MCI vs. AD) were
computed heuristically and are presented in table (4.1). Figure (4.2) shows an example of a reference
cluster.

It should be noted that the reference clusters were obtained in a 10 fold cross-validation procedure,
each cluster corresponding to one fold. After Yakushev Normalisation, data was normalised to the [0,1]
interval. This is a usual procedure, to avoid features with wide ranges dominating those with smaller
ones and prevent numerical problems [142]. Also, despite Yakushev Normalisation being less effective
for comparisons between controls and early stages of dementia and mild cognitive impairment (since
CGM reduction is less pronounced), it was also applied in these cases.

Table 4.1: The best reference cluster t threshold and radius parameters

<table>
<thead>
<tr>
<th>Modality</th>
<th>$t_{baseline}$</th>
<th>$r_{baseline}$</th>
<th>$t_{12months}$</th>
<th>$r_{12months}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. MCI</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CN vs. AD</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 4.2: Brain sagittal cut with an example of a reference cluster (black area inside the brain) obtained
from CN vs. AD comparison, located mainly in the cerebellum, which is in agreement with literature. The
white zone corresponds to a candidate to reference cluster, but discarded due to an inferior average t-value
inside.

4.4 Feature Extraction and Selection

As referred in section (4.3), each preprocessed image of patients corresponded to a number of
features in the order of millions. This posed the aforementioned problem of dimensionality (see chapter
2) and would certainly increase computation time. In addition, no useful information could be given
by those voxels outside the brain that allowed to distinguish between CN, MCI and AD patients. In
order to amend this problem, a mask was applied, a step which can be viewed as a selection procedure.
The domain of voxel intensity $V(x,y,z)$, denoted by B, can be defined by:
\[ B = \{ x, y, z \in \mathbb{N} : 1 \leq x \leq 121, 1 \leq y \leq 145, 1 \leq z \leq 121 \} \] (4.1)

The mask procedure was performed as in a previous work of Bicacro et al. [95]. It consisted of a binary mask \( M(x,y,z) \) where each position inside the brain is set to logical true. To define the voxels inside the brain, an average brain was computed using the whole database set and a threshold set at 5% of the maximum value was applied, after empirical tests (see figure 4.3). The resulting mask had approximately half a million voxels, which corresponds to a quarter of the initial feature space.

![Binary mask](image)

**Figure 4.3:** The binary mask. On the left, an example of an axial cut of a patient’s original image and the same image after applying the binary mask on the right (Slice z=61).

After this straightforward feature selection procedure, a ranking feature selection method was applied based on the t-value obtained from unpaired t-test that allowed to obtain patches from images in a probabilistic approach, as done by Tong et al. [65]. Similar strategies have been applied by Chu et al. [93] and Liu et al. [87].

Briefly, t-tests were performed on the images and each voxel was attributed a t-value. Then, a mean t-value was calculated within a surrounding cubic volume centred at each voxel. Finally, the probability of each patch was associated with the t statistic absolute value, corresponding to the intuition that a higher t-value means a more probable rejection of the null hypothesis. In practice, this corresponds to applying a mean filter to the statistic images. It should be noted that the reciprocal of the p-value could have been used instead as a measure of patch probability, but for selection purposes both statistics have similar effects (higher reciprocal corresponds to higher t-value, smaller reciprocal to smaller t-value). The t-value of each voxel was calculated according to an unpaired t-test:

\[
t_j = \frac{\mu_1 j - \mu_2 j}{s_{1,2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{4.2}
\]

where \( t_j \) is the \( j^{th} \) voxel t-value, \( \mu_1 j \) and \( \mu_2 j \) are the sample of class 1 and sample of class 2 averages (of voxel \( j \)), respectively, \( n_1 \) and \( n_2 \) their sample size, respectively, and \( s_{1,2} \) the samples common standard deviation, given by:

\[
s_{1,2} = \sqrt{\frac{(n_1 - 1)s_1^{2j} + (n_2 - 1)s_2^{2j}}{n_1 + n_2 - 2}} \tag{4.3}
\]

In theory, the number of patches \( N \) may reach the dimensionality of the image, which is extremely high. This may lead to dimensionality problems, which suggests the selection of only \( K \leq N \) patches.
These were sorted by discriminative power, in descending order of their absolute t-value. Additionally, a distance threshold calculated between the geometric centre of patches was imposed, a strategy that potentially reduces redundant information extraction [65]. Lastly, the patch size was also taken into account (see table (4.2)). It should be underlined that the approach of this thesis is to consider patches as instances and the whole brain as a bag. Although each bag could have different number of patches, preference was given to using an equal number of instances per bag.

![Image](image_url)

**Figure 4.4:** Example of a selection of 100 patches, each with size 3x3x3 and a distance threshold of 10. Not all the patches are represented, only those which intersect the cuts: (a) Sagittal plane, (b) Axial plane, (c) Coronal plane.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>[1 2 5 10 20 50 100]</td>
</tr>
<tr>
<td>$l$</td>
<td>[3 5 7 9]</td>
</tr>
<tr>
<td>$d$</td>
<td>[2 3 5 10]</td>
</tr>
</tbody>
</table>

Table 4.2: Patch selection parameters. $K$ - total number of patches; $l$ - cube edge; $d$ - distance threshold. All measures in voxel units.

The selection procedure for SVM with best features was slightly different, in that there was no selection of patches, but rather a selection of best features according to an absolute t-value ranking (see the next section). Regarding the number of patches $K$, shown in table (4.2), DD was tested only for $K \in [125]$, due to time limitations, and all others, except MILES and SVM with patches, up to 50 patches.

4.5 Classification

4.5.1 Support Vector Machines

Support Vector Machines are one of the most powerful tools in supervised machine learning, being widely used nowadays. They belong to the field of linear classifiers, such as the simple perceptron, but are much more flexible. SVM was originally developed by Vapnik in 1963 [143]. Its actual form is due to Boser, Guyon and Vapnik, who structurated the problem in a standard form [144]. One recent improvement was the introduction of a relaxing concept - the soft margin - by Vapnik and Cortes [145].

From a conceptual point of view, SVM is useful when dealing with binary problems with linearly separable sets, that is, sets which are separable by at least one hyperplane. In particular, this algorithm
finds the model of the hyperplane that maximises the margin between the two sets.

More formally, let $M$ be the number of training vectors $\vec{x}_i$, each of dimension $d$, and the corresponding class be $y_i = \pm 1$. The training set may be defined as:

$$\{\vec{x}_i, y_i\}, i = 1, ..., M, y_i \in \{0, 1\}, \vec{x}_i \in \mathbb{R}^d$$ (4.4)

Assuming the data to be linearly separable, there is a hyperplane which divides the two classes in the $\chi$ space, whose equation is:

$$\vec{w} \cdot \vec{x} + b = 0$$ (4.5)

where $\vec{w}$ is the normal to the hyperplane and $b$ is the bias term. The distance from this to the origin is $\frac{b}{||\vec{w}||}$. When $d=2$, the hyperplane corresponds to a line, as the reader can easily understand through figure 4.5.

---

**Figure 4.5:** The maximum optimal margin in a binary classification problem. Examples from two classes are separated by a hyperplane that maximises the margin between the two classes (given by $d_1 = d_2$). Instances lying on the traced line correspond to the support vectors. Adapted from [146].

It should be noted that, in more formal terms, a hyperplane parameterised by $\vec{w}$ and $b$ is said to be a separating hyperplane when there is a function $f$ over the instance space such that:

$$f(\vec{x}) = \text{sign}(\vec{w} \cdot \vec{x} + b)$$ (4.6)

Solving the SVM problem corresponds to finding the vector $\vec{w}$ and threshold parameter $b$ that are solution of the following conditions:

$$\vec{w} \cdot \vec{x}_i + b \geq 1, y_i = 1$$ (4.7)

$$\vec{w} \cdot \vec{x}_i + b \leq -1, y_i = -1$$ (4.8)

Or simply:

$$y_i(\vec{w} \cdot \vec{x}_i + b) - 1 \geq 0, \forall i$$ (4.9)
Referring to figure 4.5, \( \vec{w}.\vec{x}_i + b = 1 \) corresponds to the upper traced line, and \( \vec{w}.\vec{x}_i + b = -1 \) corresponds to the lower traced line, and \( d_1 = d_2 \) is the value of the margin, which is maximised under the SVM model, and equals \( \frac{1}{||\vec{w}||} \).

Now, in order to maximise the margin, it is possible to minimise \( ||\vec{w}|| \) instead. However, since this is equivalent to minimising \( \frac{1}{2}||\vec{w}||^2 \) and the latter is easily differentiable, finding the best hypothesis (that is, the best separating hyperplane) is accomplished through a Quadratic Programming (QP) optimisation process:

\[
\min \frac{1}{2}||\vec{w}||^2,
\]
subject to \( y_i(\vec{w}.x_i + b) - 1 \geq 0, \forall i \)

The solution of this problem is obtained with Lagrange multipliers. Consider the Lagrangian function associated with the QP optimisation problem in (4.10):

\[
L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2}||\vec{w}||^2 - \sum_{i=1}^{M} (\alpha_i (\vec{x}_i.\vec{w} + b) - 1)
\]

\[
= \frac{1}{2}||\vec{w}||^2 - \sum_{i=1}^{M} \alpha_i (\vec{x}_i.\vec{w} + b) + \sum_{i=1}^{M} \alpha_i
\]

The values of \( \vec{w}, b \) that minimise the lagrangian can be obtained by differentiation of \( L \) with respect to these variables:

\[
\nabla_{\vec{w}} L = \vec{w} - \sum_{i=1}^{M} \alpha_i y_i \vec{x}_i
\]

\[
\frac{\partial L}{\partial b} = - \sum_{i=1}^{M} \alpha_i y_i
\]

Imposing the expressions above to be zero, it follows:

\[
\vec{w} = \sum_{i=1}^{M} \alpha_i y_i \vec{x}_i
\]

\[
\sum_{i=1}^{M} \alpha_i y_i = 0
\]

Finally, differentiation of \( L \) with respect to \( \alpha \) is carried out. To do so, equation (4.14) is inserted in the expression of the Lagrangian (4.11):

\[
L(\vec{\alpha}) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{K} y_i y_j \alpha_i \alpha_j \vec{x}_i^T \vec{x}_j
\]

This equation is equivalent to the so called dual formulation (in contrast to the primal formulation, equation (4.10)).
Finally, from the Karush-Kuhn-Tucker (KKT) condition\footnote{KKT conditions are first order necessary conditions for nonlinear programming problems to be optimal, and generalise the method of Lagrange multipliers when there are inequality constraints, as is the case in SVM. The KKT condition in this situation is given by $\alpha_i (y_i (\vec{w}^T \vec{x}_i + b) - 1) = 0$, $\forall i = 1, \ldots, M$}, it is possible to conclude that for support vectors the corresponding $\alpha$ is non-negative, and otherwise zero. This result is of great importance, since the effective complexity of support vector machines depends on the number of support vectors only, not on the dimensionality of the space, which is equal to the dimension of vector $\vec{w}$. This is, perhaps, the cornerstone result of SVM, which allows for non-linear transformation of the data to infinite space dimension (e.g radial base function kernel). In practical terms, the optimal hyperplane is defined only by the support vectors (see equation (4.14)). Boser et al. [144] proved this through a very simple inequality, depending only on the number of support vectors and training instances:

$$E(\mathbb{E}_{\text{out}}) \leq \frac{E(\#\text{support vectors})}{M}$$  \hspace{1cm} (4.18)

where $E$ stands for the expected value. As a consequence, the effective capacity of SVM is usually smaller than the VC-dimension of linear classifiers, given by $d + 1$.

The explanation above presents the original SVM theory, known as hard margin model. This framework lacks the ability to classify non-linearly separable sets, such as those presented in figure 4.6. Two procedures have been developed to obviate this drawback, one based on non-linear transformation to a higher dimensional space and the other based on relaxed margin constraints.

\[
\max_{\vec{\alpha}} \vec{\alpha}^T \vec{1} - \frac{1}{2} \vec{\alpha}^T D \vec{\alpha},
\]
subject to $\vec{\alpha} \geq \vec{0}$

$$\vec{\alpha}^T \vec{y} = 0$$ \hspace{1cm} (4.17)

Figure 4.6: Two cases where the data is not linearly separable. In (a) a possible mapping is $\phi(\vec{x}) = x_1^2 + x_2^2$. In (b), applying a kernel risks to overfit the data, as clearly depicted. The introduction of a slack variable $\xi$ to account for some misclassifications may improve the generalisation.
The first approach is best known as the kernel trick. The name comes from the definition of a kernel function which implicitly defines a mapping from the original feature space to a higher dimensional one. The advantage of the kernel trick resides essentially on two aspects: an identical computational time despite of the higher dimension of the space and rendering the data in a linear separable way in the new space. In more formal terms, the kernel is defined as:

\[
K(\vec{x}_k, \vec{x}_l) = \tilde{\phi}(\vec{x}_k)\tilde{\phi}(\vec{x}_l)
\]

where \(\phi: \chi \rightarrow \mathbb{Z}\) is a non-linear transformation from the instance space to the new one. In general terms this transformation may be defined as:

\[
\tilde{\phi}(\vec{x}) = (\phi_1(\vec{x}), \phi_2(\vec{x}), ..., \phi_{\tilde{d}}(\vec{x})) = (z_1, z_2, ..., z_{\tilde{d}})
\]

where each component \(\phi_i\) of the new instance is itself a function and may have a different expression from the others and \(\tilde{d}\) is the dimension of the new space.

In general, \(\tilde{d}\) is bigger than \(d\) and this increases the risk of overfitting. However, under the machinery of SVM, dimensionality usually does not jeopardise the results except for some kernels as the RBF, as explained above (see equation (4.18)). Now, in the new space \(Z\), the weights are given by:

\[
\vec{w} = \sum_{i=1}^{M} \alpha_i y_i \tilde{\phi}(\vec{x}_i)
\]

And the linear classifier is defined by:

\[
f(\vec{x}) = \text{sign}(\sum_{i=1}^{M} \alpha_i y_i \tilde{\phi}(\vec{x}_i)\tilde{\phi}(\vec{x}) + b),
\]

\[
= \text{sign}(\sum_{i=1}^{M} \alpha_i y_i K(\vec{x}_i, \vec{x}) + b)
\]

Several types of kernels may be used, each one with its particularities. The simplest and trivial one is the linear, which corresponds to the usual inner product between to instances (\(\tilde{\phi}(\vec{x}) = \vec{x}\)) and is equivalent to the original SVM formulation. Two other renown and widely used are the polynomial \((K(\vec{x}_i, \vec{x}_j) = (\gamma \vec{x}_i^T \cdot \vec{x}_j + r)^d,\) where \(d\) is a specific degree parameter) and the radial-basis function (RBF) kernel \((K(\vec{x}_i, \vec{x}_j) = \exp(-\gamma ||\vec{x}_i - \vec{x}_j||^2)).\) Which one to use is tipically a question dependent of the problem and the specificities of the data, and there is sometimes a tradeoff. In general, the linear kernel may be more suitable when dealing with few data and many features. Otherwise, a RBF kernel might serve well [147]. A kernel must be symmetric and satisfy Mercer’s condition.

The other approach developed to deal with non-linearly separable cases was the introduction of a positive slack variable \(\xi\), which accounts for the ammount of violation of the margin, by Vapnik and Cortes [143]. The primal form in this case is given by:
\[ \begin{align*}
\min & \frac{1}{2}\|\vec{w}\|^2 + C \sum_{i=1}^{M} \xi_i \\
\text{subject to} & \quad y_i(\vec{w}.\vec{x}_i + b) \geq 1 - \xi_i, \forall i \\
\xi_i & \geq 0, \forall i
\end{align*} \]

(4.23)

where parameter \( C \) controls the trade-off between the penalty resulting from violations and the size of the margin [146]. In reality, it acts as a regularising parameter. When \( C \) tends to infinity, equation (4.23) becomes the original hard margin formulation. This optimisation problem is still convex and quadratic, and after some manipulation, its dual form is exactly the same as before, only with a slight change in one constraint:

\[ \begin{align*}
\max & \vec{\alpha}^T \vec{1} - \frac{1}{2} \vec{\alpha}^T D \vec{\alpha}, \\
\text{subject to} & \quad 0 \leq \vec{\alpha} \leq C \vec{1} \\
\vec{\alpha}^T \vec{y} & = 0
\end{align*} \]

(4.24)

In this work, the dual problem of SVM was solved with the public software LIBSVM in its version 3.17, developed by Chang and Lin [147]. Two different approaches were tested in order to compare their performance to MIL classifiers, namely SVM with the best \( k \) features, \( k \in \{100, 500, 1000, 5000, 15000\} \), and SVM with patches, where the features obtained were concatenated in \( K \times l \) dimensional vectors (see table (4.2)). Also, class unbalances were taken into account through the introduction of weights to classes. This corresponds to introducing a different parameter \( C \) to each class, instead of one (see equation (4.6)). A linear kernel was used in both approaches, and parameter \( C \in \{2^{-16}, 2^{-14}, ..., 2^{20}\} \) tuned with a search grid through nested cross-validation.

**4.5.2 CKNN**

This classifier was developed by Wang and Zucker [126] as an extension of the k-NN classifier to fit the MIL paradigm. It belongs, thus, to a lazy form of learning, where training examples are stored so that when shown a test example, these are queried [148].

As in its supervised counterpart, the labels of the \( k \) nearest neighbours of the test example are combined and the test example label is decided by majority vote. In k-NN, the Euclidean distance is a usual distance measure choice. However, under MIL, there is the necessity to define a new one, since bags are not vectorial entities, but sets of vectors. For that, Wang and Zucker applied an adapted version of Hausdorff distance.

The Hausdorff distance is defined as follows:

\[ d_H(A, B) = \max(h(A, B), h(B, A)) \]

(4.25)

where \( A = \{a_1, ..., a_m\} \) and \( B = \{b_1, ..., b_n\} \) are two sets and \( h(A, B) = \max_{a \in A, b \in B} \|a - b\| \), with \( m \) and \( n \) the cardinalities of \( A \) and \( B \), respectively (see figure (4.7)).

\[ \text{Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/} \]
Figure 4.7: Illustration of the Hausdorff distance between two bags in a bidimensional space, one with 3 instances and another with 4 instances. Informally, two sets $A$ and $B$ are within Hausdorff distance $d_H$ of each other when every point of $A$ is within distance $d_H$ of at least one point of $B$, and every point of $B$ is within distance $d_H$ of at least one point of $A$.

This metric has the drawback of being very sensitive to outliers, which can be obviated through a little modification of the term $h$ in equation (4.25):

$$h_k(A, B) = k^{th}_{a \in A} \min_{b \in B} ||a - b||$$

(4.26)

where $k^{th}$ denotes the $k^{th}$ ranked distance. When $k = m$, the distance in equation (4.25) is the same as defined above. Wang and Zucker called $h_m$ the maximal Hausdorff distance. When $k = 1$, the minimal distance of the $m$ point distances determines the value of the overall distance, since:

$$h_1(A, B) = \min_{a \in A} \min_{b \in B} ||a - b|| = \min_{b \in B} \min_{a \in A} ||a - b|| = h_1(B, A)$$

(4.27)

It follows from equation (4.27) that $H(A, B) = h_1(A, B)$, and the authors called this the minimal Hausdorff distance. It is simply the minimal distance between two sets and has the advantage of being less prone to noise. Nonetheless, the authors of CKNN have shown through the benchmark dataset Musk1 that in some cases, bags from one class were incorrectly classified by the usual majority voting process. This was due to the presence of false positive instances in positive bags, that attracted negative bags (see figure 5.8).
Figure 4.8: Four bags, two positive and two negative, are presented by their instances. Negative instances are represented by round dots, positive ones by rectangles. A misleading situation happens when \( \{P_1, P_2, N_1\} \) are used as training bags, since \( N_2 \) will be classified as positive by majority voting. In another situation, given \( \{N_1, N_2, P_1\} \) as three training bags, \( P_2 \) will be classified as negative. Source: [126].

One of the possible strategies to amend this is a citation approach. The concept of citation was defined in the context of library and information science and is related to that of reference and citer. In this context, a paper that cites a reference is said to be related to it. Similarly, a paper is also related to its citer(s). In the MIL context, a reference of an unseen bag may be translated as one of its \( R \) nearest neighbours. A citer of an unseen example may be defined as a bag that sees it as reference. In more formal terms, let \( n \) be the number of bags and \( BS = \{B_1, ..., B_n\} \) the set of all bags. Then, for an example \( B \in BS \), a bag \( B_i \in BS \setminus B \) may be ranked according to its similarity to example \( B \). Let its similarity be defined by a rank number \( \text{Rank}(B_i, B) \). The \( C \)-nearest citers of \( B \) are defined based on a rank, as follows:

\[
citers(B, C) = \{B_i | \text{Rank}(B_i, B) \leq C, B_i \in BS\}
\]  

(4.28)

Figure (4.9) shows how citers and references are assigned to bags:

<table>
<thead>
<tr>
<th></th>
<th>N=1</th>
<th>N=2</th>
<th>N=3</th>
<th>N=4</th>
<th>N=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b_1 )</td>
<td>( b_1 )</td>
<td>( b_2 )</td>
<td>( b_3 )</td>
<td>( b_4 )</td>
<td>( b_5 )</td>
</tr>
<tr>
<td>( b_2 )</td>
<td>( b_1 )</td>
<td>( b_4 )</td>
<td>( b_5 )</td>
<td>( b_2 )</td>
<td>( b_6 )</td>
</tr>
<tr>
<td>( b_3 )</td>
<td>( b_3 )</td>
<td>( b_2 )</td>
<td>( b_6 )</td>
<td>( b_4 )</td>
<td></td>
</tr>
<tr>
<td>( b_4 )</td>
<td>( b_6 )</td>
<td>( b_2 )</td>
<td>( b_3 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( b_5 )</td>
<td>( b_6 )</td>
<td>( b_4 )</td>
<td>( b_2 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.9: The nearest neighbours of 6 bags \( \{b_1, b_2, b_3, b_4, b_5, b_6\} \). \( N \) is the nearest rank number. Let \( R \) and \( C \) be both 2. Then, for bag \( b_1 \), its \( R \)-nearest references are \( b_3 \) and \( b_2 \) and its \( C \)-nearest citers are amongst \( b_2 \), \( b_3 \) and \( b_5 \). Source: [126].

Finally, to obtain the class of an unseen bag, it is somehow necessary to combine both the \( R \)-nearest references and the \( C \)-nearest citers. Assuming that among the \( R \)-nearest references there are \( R_p \) positive bags and \( R_n \) negative bags and \( C_p \) and \( C_n \) is the number of positive and negative \( C \)-nearest citers, respectively, it is possible to define \( p = R_p + C_p \) and \( n = R_n + C_n \). Quite intuitively, if \( p > n \),
then the class of the bag is predicted as positive, otherwise negative. In case of tie, the authors set the result to negative.

In this work, the search for the best C and R parameters was performed with a nested cross-validation procedure (see section (4.5.7)), with a sparse search grid where each parameter had values between 1 and 46 in steps of 15. The code for the algorithm was retrieved from http://cse.seu.edu.cn/people/zhangml/Resources.htm.

4.5.3 Diverse Density

Diverse Density was one of the first MIL techniques to be developed and the first to introduce a probabilistic approach [120], thus being one of the most influential. Its author, Maron, presented it in his master thesis, entitled *Learning from ambiguity* [113]. Although the concept behind DD is not difficult to grasp, its implementation is much more delicate, as it will soon be understood by the reader.

Maron defined the measure of DD in very simple terms as the intersection of the positive bags minus the union of the negative bags (see figure (4.10)). In here, there is not only the assumption that positive bags share some common trait that distinguishes them from negative bags but also that there is an intersection between them. In reality, one should not be so optimistic, for there is noise and sparse areas in the instance space. The way Maron found to fix this limitation was the introduction of a probabilistic approach behind the generation of bags, where the location of an instance is treated as an evidence for the location of the true concept (the point in instance space that most distinguishes positive and negative bags, also called the target concept). This corresponds to a softening of the concept of intersection.

The ground for the development of DD was the Musk problem, addressed before by Dietterich [109]. As the reader should remember, a candidate musk molecule was considered positive if it had any shape along time that might be detected by smell. This corresponds to a discrete problem in the sense that there is access only to samples of the continuous instance space and is indeed the common situation in MIL problems.

Let \( \vec{x} \in \chi \) be a point in the instance space, \( \vec{c} \in \chi \) be the target concept, \( \{B_1^+, B_2^+, \ldots, B_n^+\} \) be \( n \) positive bags and \( \{B_1^-, B_2^-, \ldots, B_m^-\} \) be \( m \) negative bags, both in the training set. Then, DD is defined as:

\[
DD(\vec{x}) = P(\vec{x} = \vec{c} | B_1^+, \ldots, B_n^+, B_1^-, \ldots, B_m^-) \tag{4.29}
\]

As noted by Foulds [114], this is a measure of the probability that a point in instance space is the true concept, given the training bags. Another observation worth referring is that diverse density is based on the number of different positive bags that have instances close to the concept, which should simultaneously have no negative instances close to it. This is the reason for the use of the word *diverse*.

Applying Bayes rule to equation 4.29, one obtains:

\[
DD(\vec{x}) = \frac{P(B_1^+, \ldots, B_n^+, B_1^-, \ldots, B_m^- | \vec{x} = \vec{c}) P(\vec{c})}{P(B_1^+, \ldots, B_n^+, B_1^-, \ldots, B_m^-)} \tag{4.30}
\]
Maron considers the prior to be constant and the denominator also to be constant with respect to $\vec{c}$. Consequently, the only quantity to be computed is the likelihood of the concept:

$$L(\vec{c}|B_1^+, ..., B_n^+, B_1^-, ..., B_m^-) = P(B_1^+, ..., B_n^+, B_1^-, ..., B_m^-|\vec{x} = \vec{c})$$  \hspace{1cm} (4.31)

Further assuming the conditional independence between bags given the true concept, the latter equation can be decomposed in:

$$L(\vec{c}|B_1^+, ..., B_n^+, B_1^-, ..., B_m^-) = \prod_{1 \leq i \leq n} P(B_i^+|\vec{x} = \vec{c}) \prod_{1 \leq j \leq m} P(B_j^-|\vec{x} = \vec{c})$$ \hspace{1cm} (4.32)

As stated in [113], there is no easy way of estimating the bags’ generative model $P(B_i|\vec{x} = \vec{c})$. It is easier to obtain a general estimator for $P(\vec{x} = \vec{c}|B_i)$ since one uses instead the instances as pieces of evidence. Hence, applying once again Bayes rule in 4.32, it becomes:

$$L(\vec{c}|B_1^+, ..., B_n^+, B_1^-, ..., B_m^-) = \prod_{1 \leq i \leq n} \frac{P(\vec{x} = \vec{c}|B_i^+)}{P(\vec{c})} P(B_i^+) \prod_{1 \leq j \leq m} \frac{P(\vec{x} = \vec{c}|B_j^-)}{P(\vec{c})} P(B_j^-)$$ \hspace{1cm} (4.33)

Inserting this expression into equation 4.30 and considering a constant prior and no dependence over probabilities which are not function of $\vec{c}$, one obtains:

$$DD(\vec{x}) \propto \prod_{1 \leq i \leq n} P(\vec{x} = \vec{c}|B_i^+) \prod_{1 \leq j \leq m} P(\vec{x} = \vec{c}|B_j^-)$$ \hspace{1cm} (4.34)

Now, one has just to compute the maximum likelihood:

$$\arg \max_{\vec{x} \in \chi} \prod_{1 \leq i \leq n} P(\vec{x} = \vec{c}|B_i^+) \prod_{1 \leq j \leq m} P(\vec{x} = \vec{c}|B_j^-)$$ \hspace{1cm} (4.35)

![Figure 4.10: Four bags (three positive and one negative) are depicted in a sampled bidimensional space. Each represented point is an instance. Point A corresponds to an high density of instances from positive bags and far from negative instances, which implies being a high diverse density point. Zone B is an area of high density. Nonetheless, it is close to negative instances, which forbids it from being a candidate to the target concept that maximises diverse density. Source: [113].](image)

Maron defined the probabilities above according to the most likely-cause model, which selects for each bag the instance with the highest probability of being positive:
\[ P(\vec{x} = \vec{c}B_i^+) = \max_j P(\vec{B}_{ij}^+ \in \vec{c}) \]
\[ P(\vec{x} = \vec{c}B_i^-) = (1 - \max_j P(\vec{B}_{ij}^- \in \vec{c})) \]

where \( \vec{B}_{ij} \) designates instance \( j \) belonging to bag \( i \).

Finally, it is only necessary to define an expression to compute the probability that an instance belongs to a given concept:

\[ P(\vec{B}_{ij} \in \vec{c}) = \exp\left(- \sum_{1 \leq l \leq k} (B_{ijl} - c_l)^2 \right) \]

\[ = \exp(-||\vec{B}_{ij} - \vec{c}||^2) \]

where \( \vec{B}_{ijl} \) designates feature \( l \) of instance \( j \) belonging to bag \( i \).

The expression above defines what Maron termed single point concept class, which assumes that every concept corresponds to a single point in the instance space and each positive bag contains at least one instance that is equal to the true concept corrupted by some gaussian noise. Notwithstanding, the author developed a more complete approach, the single point-and-scaling concept class (see figure (4.12)):

\[ P(\vec{B}_{ij} \in \vec{c}) = \exp(- \sum_{1 \leq l \leq k} (w_l B_{ijl} - c_l)^2) \]

\[ = \exp(-||\vec{B}_{ij} - \vec{c}||^2) \]

where now \( ||\vec{B}_{ij} - \vec{c}||^2 = \sum_{1 \leq l \leq k} w_l^2 (B_{ijl} - c_l)^2 \).

This is nothing more than learning the best weights for features and is, in fact, a form of regularisation. The reason for its implementation lies on the existence of many irrelevant features in real situations, which leads to the need for some sort of feature weighting \[113\]. Attributes with high scale factor are considered more important, those with smaller scale factor, less important. In a geometric analysis, one can interpret the scaling vector as a space mold. Negative bags tend to increase the space along its dimensions, and positive bags the contrary (see figure (4.12)). Under this formulation, a concept is defined by the tuple \( (\vec{c}, \vec{w}) \). However, there is now the double number of parameters to specify the true concept, which may also lead to overfitting.

The last question which arises is how to find in an efficient way the concept in the instance space that maximises DD. This is in general a difficult global optimisation problem, for the concept space is at least as big as the feature space, continuous and high dimensional \[113\]. Maron proposed the so called maxDD algorithm, which performs a gradient ascent based optimisation in the concept space, searching for the point which is closest to instances from positive bags and far from all negative instances. One of the drawbacks of this optimisation process is the possibility of getting stuck in local maxima. One good heuristic that diminishes this risk is to perform multiple optimisations, each started in a different random point or a priori good points, whenever known. Fortunately, MIL allows for the latter case, and a satisfactory approach is starting gradient ascent in instances belonging to positive bags. Despite
not being guaranteed to reach the global maximum, this a good heuristic, with successful results in practice, even if carrying a high computation cost when dealing with a large number of instances in positive bags [113]. This is the general procedure when dealing with single point concept (see figure (4.11)). With respect to the point-and-scaling concept formulation (see figure (4.12)), there are now twice the parameters to find and this heuristic may also be applied to find the best scaling vector. Only one does not know a priori good initialisers in the scaling space. Maron opted to arbitrarily start this vector ($\vec{w}$) with all entries equal to one.

In terms of complexity, DD increases the computation time with the total number of instances in all bags, being $O(N)$ both in time and space [113]. Besides, the optimisation procedure may be slow, particularly when dealing with single point and scaling concept. Despite this time drawback, it was used in this thesis.

Finally, in order to classify a bag, one selects the distance threshold $d$ (which is equivalent to a probability threshold, see equation 4.38) that maximises the accuracy on the training set, according to a nested cross-validation procedure, a technique explained further below in this chapter. More specifically, the probability threshold was tuned in a search-grid between 0.1 and 1 in steps of 0.05. A bag is positive if at least one of its instances is within distance $d$ of the true concept, which makes DD an instance-level classifier which uses the standard MI assumption.

The code for this algorithm was retrieved online from http://cse.seu.edu.cn/people/zhangml/Resources.htm, developed by Zhou and Zhang under the scope of their article about ensemble methods on MIL [125].

A final word about the extraction of patches for Diverse Density is necessary. In fact, selected patches, as defined in section (4.4), are not adequate for DD. A close look in the way the algorithm is built shows that there must be a correspondence between instance features (in the same way SVM needs it), that is, feature $j$ of a given instance must correspond to feature $j$ of all other instances. Patches from brain images are selected in different areas of the brain, thus preventing any equivalence.
between features. This was circumvented through the use of histograms of patches. In each cross-validation fold, each histogram had as many bins as needed between the minimum intensity voxel value and the maximum intensity voxel value of all training images, in steps of 0.05 (usually around 20 bins). Besides, only up to 5 patches were extracted, due to computation time limitations.

![Diagram](image)

**Figure 4.12:** The effect of weights on diverse density. When introduced, these allow a better maximisation process. Source: [113].

### 4.5.4 MILES

MILES was developed by Chen et al. [133] as a wrapper classifier, where the 1-norm SVM is used to classify and select the most important features.

In broad terms, MILES has the capability of identifying the most relevant instances and selecting those through the definition of a similarity measure between bags and instances. This is accomplished through an innovative way of bag representation, in an embedded space where each feature is associated to an instance, and thus feature selection becomes equivalent to instance selection. The embedding step makes MILES belong to vocabulary based methods.

Conceptually, MILES owes much to DD. In fact, it extends this classifier in that multiple concepts are allowed, belonging to positive or negative examples. In DD, there was a rather restrictive assumption of a single concept belonging to positive examples.

Let \( \bar{c} \in C \) be a concept belonging to concept class \( C \). The likelihood of a bag given this concept is defined as:

\[
L(B_i|\bar{c}) = P(\bar{c}|B_i) \tag{4.39}
\]

If \( C \) is a countable set, then it is possible to define:

\[
[L(B_i|\bar{c}_1), L(B_i|\bar{c}_2), ..., L(B_i|\bar{c}_j), ...] = [P(\bar{c}_1|B_i), P(\bar{c}_2|B_i), ..., P(\bar{c}_j|B_i), ...] \tag{4.40}
\]

where the member to the right represents a bag under MILES framework. It remains the question how to calculate the posterior probability of a concept. Chen et al. opted for the use of the most-likely cause estimator [113] independently of the bag label.
\[ P(\tilde{c}_k | B_i) \propto s(\tilde{c}_k, B_i) = \max_j \exp\left(-\frac{||\tilde{x}_{ij} - \tilde{c}_k||^2}{\sigma^2}\right) \] (4.41)

where \( s \) represents the similarity between concept \( \tilde{c}_k \) and bag \( B_i \), \( \tilde{x}_{ij} \) represents an instance belonging to bag \( B_i \) and \( \sigma \) a predefined factor. A bag may now be mapped into a new space, defined as:

\[ \tilde{M}(B_i) = [s(\tilde{c}_1, B_i), s(\tilde{c}_2, B_i), ..., s(\tilde{c}_n, B_i)] \] (4.42)

where \( n \) is the dimension of the new vector, equal to the number of instances from all training examples (concepts). In fact, each concept corresponds to an instance belonging to training bags.

Note that each feature of vector \( M \) may be interpreted as a similarity between a bag and a concept. The authors of MILES consider that the target concept points may be approximated by instances in training bags, irrespective of their label, which leads to a symmetric approach, and thus, vector \( M \) summarises the similarities between a bag and all training instances [111]. Desirably, if \( \tilde{c}_k \) shows a bigger similarity to positive bags and low similarity to negative bags, feature \( s(\tilde{c}_k, B_i) \) might be useful in separating the two classes of bags [133].

Figure 4.13: The embedding procedure. Instances from positive bags (red circumferences) and negative bags (in blue triangles) lie on a bi-dimensional space, where there is no clear separation between them (a). Assuming there are three concepts, the embedding maps instances to a three-dimensional space, where it is much easier to distinguish bags. In the embedded space, each bag is defined by three features. Source: [133].

MILES was originally built using 1-norm SVM, a variant of SVM differing in the regulariser, which is in this case the Manhattan norm of the weight vector. This regulariser causes the weight vector to have more null entries (thus the alternative name of sparse SVM). Moreover, it renders the optimisation problem a linear programming one, instead of quadratic, which allows an improvement in its resolution speed [114]. However, it should be noted that any other single-instance classifier might be used inside MILES.

It is easy to see, from equation (4.42), that features in the new space correspond to instances. Consequently, weights learnt by SVM are, in fact, instance weights. Despite defining weights for the training instances, there is not a clear weight function over instance space, as Foulds noted [114].
results from the way the similarity $s(\vec{c}, B)$ between a training instance and a bag is defined, which only accounts for the closest instance inside the bag (see equation (4.41)). All other instances are neglected when classification occurs.

Considering the simplest possible case, where a bag only has one instance $\vec{x}$ and there is only one training instance $\vec{c}_k$, instance $\vec{x}$ final weight is defined by:

$$w(\vec{x}) = w^*[k] \exp \left(- \frac{||\vec{x} - \vec{c}_k||^2}{\sigma^2} \right)$$

(4.43)

where $w^*$ is the weight output by SVM. It follows that the classification of the bag is given by:

$$y(B) = sign(\vec{w}.\vec{M}(B) + b)$$

$$= sign \left( \sum_k (w^*k \exp \left(- \frac{||x - \vec{c}_k||^2}{\sigma^2} \right) + b^* \right)$$

(4.44)

where $w_f(x)$ is the influence the instance $x$ belonging to bag $B$ has in its classification. Being $w^*$, $b^*$, $\vec{c}_k$ and $\sigma$ all constants, $w_f$ is indeed a function over the instance space. However, in the more common situation of having many instances per bag, the influence of an instance on the bag label depends inevitably on the other instances, due to the max operator in equation (4.41). As a result, $w_f(x) : \chi \times B \rightarrow \mathbb{R}$ depends both on the instance space and bag space. In terms of complexity, MILES is $O(nm)$, where $m$ is the cardinality of each bag and $n$ is the total number of concepts.

One of the interesting aspects of MILES is the possibility of feature classification and instance classification. Since MILES is a classifier whose input features are defined by training instances, weights attributed to each feature will carry information about its respective instance. The authors of MILES divided features into positive ($w_k > 0$), negative ($w_k < 0$) and void features ($w_k = 0$). Remarkably, in image analysis, every positive feature was observed to be defined by an image patch from a positive image and every negative feature by a patch from a negative image, without imposing any constraints on the classification problem. Chen et al. expressed, nonetheless, a certain caution in the interpretation of these results, pointing out the existence of false positive features [133]. As it can be seen, feature classification is in fact concept classification (training instance classification). Another relevant aspect related to feature weights is the sparsity of MILES when using 1-norm SVM, which was reported to be around 98 percent. Finally, MILES appears to be robust to labeling noise, that is, mislabeling of bags. What concerns instance classification, this may be achieved by:

$$y_{ij^*} = \sum_{k \in I_{j^*}} \frac{w_k s(\vec{c}_k, B_{ij^*})}{m_k}$$

(4.45)

where $y_{ij}$ is the label of instance $j$ of bag $i$, $I_{j^*}$ is the set of features whose closest instance in bag $i$ is $j^*$, $w_k$ is the weight of each feature $k$ and $m_k$ is the number of closest instances to feature $k$ in bag $i$. It should be noted that not all instances belonging to a bag may be classified. In fact, those which are not the closest to concepts are not assigned any label given that they do not have any weight defined, as discussed before.
In this thesis, MILES was implemented with the usual 2-norm SVM classifier, using a linear Kernel. As noted by Foulds, the implementation of 1-norm SVM has no significative improvement effect in the performance of MILES in general. In particular, the sparsity of this method did not translate into consistently superior classification accuracy [114]. Parameters $\sigma^2$ and $C$ were obtained through nested cross-validation in a search grid, for $\{2^6, 2^8, \ldots, 2^{14}\}$ and $\{2^{10}, 2^{12}, \ldots, 2^{22}\}$ values, respectively, which were found heuristically to be the most appropriate. Figure (4.14) shows the pseudocode of the algorithm.

\[
\begin{align*}
D &= \text{the set of training bags} \\
C &= \text{all instances in the bags in } D \\
L &= \text{a single-instance base learner} \\
\sigma &= \text{the scaling factor, a parameter to the algorithm} \\
\end{align*}
\]

\[
MILES_{\text{transform}}(B) \\
\text{for (every instance } x^k \text{ in } C \text{ do} \\
\quad d = \min_j \|x_j - x^k\| \\
\quad \text{the } k\text{th element of } m(B) \text{ is } s(x^k, B) = e^{-\frac{d^2}{\sigma^2}} \\
\text{return } m(B)
\]

\[
\begin{align*}
\text{train}(D) \\
F &= \text{an empty set of instances} \\
\text{for (every bag } B_i = \{x_j : j = 1, \ldots, n_i \text{ in } D \text{ do} \\
\quad t = MILES_{\text{transform}}(B_i) \\
\quad t.setClassLabel(B_i, getClassLabel()) \\
\quad F = F \cup \{t\} \\
L.train(F) / / \text{Can optionally perform feature selection here also}
\end{align*}
\]

\[
\begin{align*}
\text{classify}(B), B = \{x_j : j = 1, \ldots, n_i \text{ a test bag} \\
\bar{t} = MILES_{\text{transform}}(B) \\
\text{return } L.classify(\bar{t})
\end{align*}
\]

Figure 4.14: MILES pseudocode. Source: [114].

4.5.5 YARDS

YARDS was developed by Foulds [114] as a modified version of MILES. Its main aim was to render the learnt instance weights an effective function over the instance space. As explained above, MILES implemented $w(x) : \chi \times B \rightarrow \mathbb{R}$, which was not a pure function of instance space.

As MILES, YARDS maps each bag into an instance-based feature space where each feature represents the similarity between this bag and a training instance and then applies some supervised classifier for classification. For the first, this similarity was given by the most-likely cause estimator (see (4.41)), which accounted only for the closest instance inside a bag. YARDS excludes this dependency by replacing the max operator by a sum in the equation. For the rest, everything is identical:

\[
P(\tilde{c}_k | B_i) \propto s(\tilde{c}_k, B_i) = \sum_j \exp(-\frac{\|x_{ij} - \tilde{c}_k\|^2}{\sigma^2}) \quad (4.46)
\]

In order to prove that YARDS produces indeed a weight function over the instance space, and using the same notation as in MILES, let $\bar{M}(B)$ be the embedded vector that represents bag $B$ after the mapping, then:
\[ y(B) = \text{sign}(\vec{w}, \vec{M} + b) \]
\[ = \text{sign}\left(\sum_k w_k M_k + b\right) \]
\[ = \text{sign}\left(\sum_k w_k s(\vec{c}_k, B) + b\right) \]
\[ = \text{sign}\left(\sum_k w_k \sum_j \exp\left(-\frac{||\vec{B}_j - \vec{c}_k||^2}{\sigma^2}\right) + b\right) \]
\[ = \text{sign}\left(\sum_k \sum_j \exp\left(-\frac{||\vec{B}_j - \vec{c}_k||^2}{\sigma^2}\right) w_k + b\right) \]
\[ = \text{sign}\left(\sum_j \sum_k \exp\left(-\frac{||\vec{B}_j - \vec{c}_k||^2}{\sigma^2}\right) w_k + b\right) \]

Now, defining a constant \( d \) which values one whenever the exponential in the latter expression is positive and minus one otherwise, the exponential becomes:

\[ \sum_k \exp\left(-\frac{||\vec{B}_j - \vec{c}_k||^2}{\sigma^2}\right) w_k \]
\[ = |\sum_k \exp\left(-\frac{||\vec{B}_j - \vec{c}_k||^2}{\sigma^2}\right) w_k|d \]
\[ = w(\vec{B}_j)d \]

Finally, one can derive the expression for the label of bag \( B \):

\[ y(B) = \text{sign}\left(\sum_j w(\vec{B}_j) + b\right) \]

As the reader can easily see, \( w(\vec{x}) : \chi \to \mathbb{R}^+ \) is indeed a function over the instance space. Foulds gives an easy interpretation to the weights, in which the weight of an instance depends on the sum of a set of gaussian functions related to the distance from the distance to each training instance, the center point of each gaussian. Besides, as MILES, it is possible to label concepts.

As in Foulds thesis, YARDS was implemented with the 2-norm SVM classifier, using a linear Kernel. Parameters \( \sigma^2 \) and \( C \) were obtained through nested cross-validation in a search grid, for \( \{2^6, 2^8, ..., 2^{14}\} \) and \( \{2^{10}, 2^{12}, 2^{22}\} \) values, respectively. What concerns the number of patches \( K \), its highest value was 50, due to time limitations. Figure (4.15) depicts the pseudocode of YARDS.
BARTMIP was developed by Zhang and Zhou [132] in 2008, presenting a novel approach in MIL. As a matter of fact, it was based in the first multi-instance clustering method, BAMIC (BAg-level Multi-Instance Clustering), presented by the authors in the same paper.

Conceptually, the idea of BARTMIP is similar to MILES and YARDS. Its vocabulary, however, is obtained in an unsupervised way and its words are bags (more precisely, classes of bags), not instances, and a new inter-bag distance is employed. As a vocabulary method, each bag is first mapped to a embedded space.

In order that the reader has a glimpse of what is behind this algorithm, BAMIC is first briefly presented. This unsupervised approach is nothing more than an adjustment of the well known k-medoids algorithm to the framework of MIL, where there are bags instead of instances as main entities. Concretely, it tries to partition the unlabeled training bags into k disjoint clusters whose centroids are members of the data set (bags), not an aleatory point in space, hence being medoids.

Let \( \chi \) denote the domain of instances and \( T = X_1, X_2, ..., X_N \) be the training set, where \( X_i \in \chi \). BAMIC assembles bags in T into k disjoint clusters \( G_i, 1 \leq i \leq k \). The distance used is the average Hausdorff distance between two bags A and B:

\[
d_{aveH}(A, B) = \frac{\sum_{a \in A} \min_{b \in B} ||a - b|| + \sum_{b \in B} \min_{a \in A} ||b - a||}{|A| + |B|} \quad (4.50)
\]

This distance averages the distances between each instance in one bag and its nearest instance in the other bag, thus taking into account more information than the maximal Hausdorff distance or the
minimal Hausdorff distance. The medoid of each cluster may now be defined as the bag with the
minimum average distance to the other members of the respective cluster.

\[ [\text{Groups, Medoids}] = \text{BAMIC}(U, k, \text{Bag}_\text{dist}) \]

**Inputs:**
- \( U \) – unlabeled multi-instance training set \( \{X_1, X_2, \ldots, X_N\} \) \( X_i \subseteq \mathcal{X} \)
- \( k \) – number of clustered groups
- \( \text{Bag}_\text{dist} \) – distance metric used to calculate distances between bags, which could take the form of maximal, minimal or average Hausdorff distance in this paper

**Outputs:**
- \( \text{Groups} \) – clustering outputs \( \{G_1, G_2, \ldots, G_k\} \) \( \bigcup_{i=1}^{k} G_i = U, G_i \cap_{i \neq j} G_j = \emptyset \)
- \( \text{Medoids} \) – Medoids of clustered groups

**Process:**
1. Randomly select \( k \) training bags as the initial medoids \( C_j \) (\( 1 \leq j \leq k \));
2. repeat
3. for \( j \in \{1, 2, \ldots, k\} \) do
4. \( G_j = \{C_j\} \);
5. for \( i \in \{1, 2, \ldots, N\} \) do
6. \( \text{index} = \arg \min_{j \in \{1, 2, \ldots, k\}} \text{Bag}_\text{dist}(X_i, C_j) \);
7. \( G_{\text{index}} = G_{\text{index}} \cup \{X_i\} \);
8. for \( j \in \{1, 2, \ldots, k\} \) do
9. \( C_j = \arg \min_{A \in G_j} \left( \sum_{B \in G_i} \text{Bag}_\text{dist}(A, B) / |G_j| \right) \);
10. until [the clustering results do not change];
11. \( \text{Groups} = \{G_j\}_{1 \leq j \leq k} \); \( \text{Medoids} = \{C_j\}_{1 \leq j \leq k} \);

Figure 4.16: BAMIC pseudocode. Source: [132].

In this work, BAMIC was implemented as an adaptation of the Partitioning Around Medoids (PAM). It is essentially constituted of two phases, nominated the build phase and swap phase, respectively. In the first one, \( k \) examples are randomly selected as the initial medoids. Then, based on a dissimilarity matrix computed with the average Hausdorff distance, each example is assigned its closest medoid. Finally, the swap phase looks for examples inside each cluster that might lower a given cost function (the sum of average Hausdorff distance to a medoid in this case). If such example is found, then new clusters are rebuilt. This process is repeated until no change happens in the medoids set (see the pseudocode in figure (4.16)).

In terms of quality of the clustering algorithm, the authors suggest the use of two criteria: average purity and average entropy. Informally, these are external measures which are related to how representative of true classes the clusters are. Let each instance have a weight whose value is inversely proportional to the size of its bag \( \left( \frac{1}{|B_j|} \right) \) and \( W_j = W_j^0 + W_j^1 \) be the sum of weights of all instances in cluster \( G_j \), where the superscripts indicate the two possible labels for bags. The average purity of clusters is given by:

\[
\text{avpurity}(\{G_1, G_2, \ldots, G_k\}) = \frac{1}{k} \sum_{j=1}^{k} \frac{W_j}{N} \text{purity}(G_j) = \frac{1}{k} \sum_{j=1}^{k} \frac{W_j}{N} \max \left\{ \frac{W_j^0}{W_j}, \frac{W_j^1}{W_j} \right\}
\]

(4.51)

The purity of a cluster can be perceived as a measure of its consistency or homogeneity in terms of class labelling. In the best scenario, its value equals one, and zero otherwise. Finally, average entropy of clusters is defined as:
\[ \text{aventropy}(\{G_1, G_2, \ldots, G_k\}) = \sum_{j=1}^{k} \frac{W_j}{N} \text{entropy}(G_j) = \sum_{j=1}^{k} \frac{W_j}{N} \left( \sum_{i \in \{0,1\}} \frac{W^i_j}{W_j} \log \frac{W^i_j}{W_j} \right) \] (4.52)

Entropy can be thought of as the information needed to correctly classify all the bags in clusters and should ideally be zero.

After the clustering step, BARTMIP is finally called, under the assumption that the k groups obtained encode some information about the different classes of bags. BARTMIP treats each bag as a k-dimensional feature vector whose i-th feature corresponds to the distance between the bag and the centroid of the i-th cluster and then applies some conventional supervised learner to the data (pseudocode in figure (4.5.6)). In this work, as in the original paper, 2-norm SVM with linear kernel is the elected supervised classifier, with C parameter chosen by nested cross-validation from \( \{2^{-10}, 2^{-8}, \ldots, 2^{20}\} \). Also, the maximum number K of patches per bag was 50, due to time limitations.

In more formal terms, given k clusters \( \{G_1, G_2, \ldots, G_k\} \) with respective medoids \( \{C_1, C_2, \ldots, C_k\} \), a test bag will be constructed as \( B = (B_1, B_2, \ldots, B_k) \), with each feature \( B_i = d_{\text{aveH}}(B, C_i), i = 1, \ldots, k \).

Another important subject is the number of clusters k. This could be derived through validation, but different values were tried instead, more precisely 2, 10, 30 and 50, although in [132] the authors suggested that this parameter has not a significant impact on classification accuracy.

---

**Figure 4.17:** BARTMIP pseudocode. Source: [132].

### 4.5.7 Nested Cross-Validation

Validation is one of the most important procedures in machine learning, the main purpose being model selection in classification in such a way that overfitting is reduced and the classifier error is
minimised. Indeed, in many situations there is the necessity to choose some parameters, such as those related to SVM kernels or the regulariser parameter.

In broad terms, when doing validation, the data set is split into a training set and a validation (or, also misleadingly termed, test) set, which are separate from the test set. The first set is used to learn the model and the validation one to select the best model. Finally, the test set evaluates the performance of the chosen model. It is important to ensure all the three mentioned sets are mutually separated, since this would lead to a highly biased and optimistic solution, generally providing low out-of-sample error (re-substitution error) [149]. Also, it should be noted that the validation error is always optimistic, although approximating in a better way the test error.

Regarding validation methods, there are different modalities, each with its own advantages. In this work, it was opted to use k-fold cross-validation (CV). This approach is among those with less bias introduced, guaranteeing at the same time a good generalisation error when few data is at hand. This is indeed the case of the provided ADIAR data, where a splitting into train and validation sets would lead to very small samples.

Specifically, k-fold cross validation was used. In brief, the data set is randomly partitioned into k disjoint sets, from which k-1 sets are used for training and 1 for validation. This procedure is repeated k times, so that each partition is selected exactly once as validation set. One common hitch in the literature is to use this methodology to select the best parameter, which violates the principle that all selection processes, i.e., feature selection or parameter selection, must be done on the data not left out [149]. In fact, with a closer glance, the reader may conclude that all the validation data, which is simultaneously used as test, is used in parameter tunning. This introduces, in fact, a bias.

Varma and Simon [149] showed that a slight variation of k-fold CV might obtain results with less bias. The authors suggested what they called a nested CV procedure, which has 2 loops of cross-validation. In the first, data is split into k sets, as before, but now one is treated as a test set which is left untouched until necessary for classification. The other k-1 enter in a new CV loop, being split in k’ sets. One of them is used as validation set, and the remaining as training set. In short, the inner CV loop is used as model selection step, repeated k’ times, and the outer one as error evaluation, repeated k times. The CV error is the average over the k outer loops. Figure (4.5.7) shows graphically how nested CV works.
Figure 4.18: Partitioning performed in the nested cross-validation procedure. Within the CV procedure, the dataset is partitioned into k folds. A different fold is used as the test set (in red), in each iteration, and the remaining ones enter several CV procedures, a process known as nested cross-validation. Within each CV, the dataset is partitioned into k' folds and, in each iteration, one is used as the validation set (red fold in last row), while the others constitute the training set (blue folds in last row). Parameter selection is performed based on the classifications attained in the validation sets, while performance assessment is based on the classification in the test sets.

4.6 Longitudinal Analysis

All the algorithms and proceedings presented above were applied to cross-sectional data, that is, baseline and 12 months data separately. Besides this analysis, it was tested a longitudinal one, assembling the two time instants.

This analysis was tested with SVM with patches and MILES and shared all preprocessing steps explained before, the only difference lying on the way patches were selected, so that both time instants could be considered. Hence, an average patch t-value sorting was computed for the two instants and then patches were selected from this sorted list. Some bags would have only patches from one instant, others from both.

4.7 Simulation of non-registered images

Image registration plays an important role in many applications, namely in medical imaging. The process is often required such that non-aligned and distorted images may be compared in a common space, the registred space. Without it, voxelwise comparison becomes almost impossible, and segmentation techniques are then required.

Despite its unquestionable value, image registration methods may become extremely time expensive when there are rotation and scaling differences [150]. This is indeed the case of ADNI data, where, besides having different scanner origins, images have intra and interpatient differences.

Bearing this drawback in mind, it was hypothesised that the MIL framework could possibly handle to some extent non-registered images. To accomplish this, rather then unregistering images, it was opted for a simpler and more straightforward way that could capture the essence of raw images, which is
the inexistence of spatial correspondence between them. Hence, an aleatory patch selection procedure was implemented, with the only restriction that at least one of the patches of each patient should belong to a discriminative area, so that MIL would apply. The patches belonging to these areas were selected according to a t statistic sorting, as explained before, but with randomness, therefore being randomly selected from the best ones. The remaining patches were selected aleatorily from the whole brain. These procedures allowed for repetition of patches amongst the patients, thus enabling some, even if low, spatial correspondence, and were implemented in each run of CV for the training set and test set.

MILES was the algorithm chosen to assess the performance of this method, in the baseline data only and for patches of size 3x3x3 with a distance threshold set to 3, for the CN vs. AD modality.
# Experimental Results

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Introduction</td>
<td>66</td>
</tr>
<tr>
<td>5.2 Methodology discussion</td>
<td>66</td>
</tr>
<tr>
<td>5.3 Neuroimaging data</td>
<td>67</td>
</tr>
<tr>
<td>5.4 Results</td>
<td>68</td>
</tr>
</tbody>
</table>
5.1 Introduction

This chapter presents the reader with a detailed analysis of the results obtained in the current work. There is first a brief discussion on the methodology followed to analyse the outcome and the pros and cons of some state of the art approaches. The remainder of the chapter follows the structure of the thesis: first, Yakushev Normalisation effects are looked into. Secondly, there is an overview of feature selection procedure in terms of regional patch counts based on regions defined by an expert doctor. Thirdly, each algorithm performance is analysed both at the baseline and 12 months. Longitudinal results will also be covered, for SVM with patches and MILES. In terms of statistics, balanced accuracy, sensitivity, specificity and AUC of ROC curves constitute the core data of this chapter. A comparison between the different algorithms is then performed with the aid of these measures. Finally, the simulation of non-registered images with MILES is assessed.

5.2 Methodology discussion

In general, comparing the performance of different classifiers is a difficult problem. A number of problems usually found in the literature has been highlighted by Japkowicz and Shah in their book about this topic [151]. Many of them arise from what these authors termed the de facto culture, which somehow instils some methodologies in disregard for others. Some paradigmatic cases are the use of cross-validation, t-tests and the choice of significance values, which happen to be almost negligently used. Also, some metrics widely used depend on data class imbalances. This is the case of accuracy. Other problems arise when comparing multiple classifiers in multiple data domains, where some aggregation measures are usually applied, such as the average, or a win/tie/loss approach, which have drawbacks.

When little data is available, as in the present thesis, the comparison becomes even more complex, a topic which has already been addressed by some authors [152, 153]. Small data means that it must be used multiple times (usually splitting it in smaller sets), rendering independence assumptions obsolete [152]. This is indeed the case of k-fold cross validation. When \( k = 10 \), for instance, each fold training set shares 80% of the examples with the others. One of the consequences is the increase of type I error, that is, the rejection of the null hypothesis when it is true with higher probability. Besides, many algorithms suffer from low replicability, which implies that the same data set and hypothesis set lead to different results.

Different approaches for comparison of classifiers have been used in the past, such as using paired t-test with resampling or k-fold cross validation. The first was reported to have extremely high type I error by different authors [152, 153]. The second was considered by Dietterich as a viable option, showing a high statistical power, but at the expense of a somehow high type I error.

An extensive analysis was performed by Bouckaert [152] where the author suggested some alternatives, namely one way to augment the replicability in k-fold cross validation by increasing the number of runs and calibrating the number of degrees of freedom of the t-test, which in some way accounts for the dependency of each fold training set, a procedure that showed reasonable type I error. Diettrich
also suggested the use of McNemar’s test, but emphasising two drawbacks: its inability to measure variability due to the choice of only one training set and the fact that this set is smaller than the whole data set, which implies the assumption that the algorithm performance changes smoothly with the size of training data [153]. Salzberg analysed common mistakes in classification comparisons and suggested the use of k-fold cross-validation with one run and then resorting to a binomial test or a McNemar test [154].

In this thesis, the option was the use of a paired t-test with 10 runs 10-fold cross validation for intra and inter-classifier comparison, where all the 100 individual balanced accuracies are used to estimate the mean and variance, with 10 degrees of freedom, as suggested by Bouckaert. This approach was termed as *use all data*, given that each fold is used as a realisation of the statistical variable (balanced accuracy). It shows high replicability, low type I error and is simple [152].

With regard to balanced accuracies (average of sensitivity and specificity), these guarantee robustness when imbalanced data is present. It should be noted that the t-test is based on the normality assumption, the randomness of samples and equal variance of the populations. The first assumption requires that samples are driven from a normally distributed population or to have at least 30 examples, so that the central limit applies. In what concerns the second condition, this requires a random selection for the testing set. The last assumption is self-explanative and is perhaps the most contestable.

Besides, whenever possible, ROC curves were employed. This method is one of the tools most used in medical diagnosis (specially imagiology and radiology) field to assess the performance of a classifier and provide a straightforward way to compare different methods, through the Area Under the Curve. Furthermore, it is robust to skewed data, which means it takes the class distribution into account.

As a final remark, it should be noted that not all the results obtained in this thesis are presented in this chapter. In fact, this would require the presentation of at least 128 (the number of patch parameter combinations) results per algorithm and would render the reading a lengthy and tedious process. Rather than this, the aim of this chapter was to present expressive results in a summarised way. Every AUC, balanced accuracy, sensitivity or specificity referred to corresponds to the average of results over 10 runs and 10 folds, except otherwise stated. When referring to results inside the text, these will be presented in the form (balanced accuracy (%), standard deviation (%)).

### 5.3 Neuroimaging data

All the data were collected by the Alzheimer’s Disease Neuroimaging Initiative, a large multisite study that focuses on the collection and analysis of different types of neuroimages, namely PET and MRI images. ADNI is a large ongoing multicenter study established in 2004 with the purpose of developing genetic, clinical, imaging and biochemical biomarkers for early diagnosis of AD. One of the major accomplishments of ADNI has been the development of standardised methods for imaging techniques, such as PET [42].

The dataset used in this study contained FDG-PET images acquired from 54 patients suffering from Alzheimer’s disease, 128 with Mild Cognitive Impairment and 74 Normal Controls, at two instants:
the initial visit (baseline) and the follow-up visit at 12 months. From the pool of data provided by ADNI, only those corresponding to the intersection of both time instants and satisfying some CDR restrictions were considered: 0 for normal controls, 0.5 for MCI patients and 0.5 or higher for those suffering from AD (see table 5.1).

Table 5.1: Clinical and demographic description of the data set used, which corresponds to the intersection of data from the baseline and 12 months. Data includes gender, age and clinical information.

<table>
<thead>
<tr>
<th>Group</th>
<th>CN</th>
<th>MCI</th>
<th>AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>74</td>
<td>128</td>
<td>54</td>
</tr>
<tr>
<td>Age</td>
<td>76.1 ± 4.7</td>
<td>74.9 ± 7.3</td>
<td>76.8 ± 6.6</td>
</tr>
<tr>
<td>Sex(% of males)</td>
<td>64.9</td>
<td>65.6</td>
<td>51.9</td>
</tr>
<tr>
<td>MMSE (baseline)</td>
<td>29.1 ± 1.0</td>
<td>27.2 ± 1.6</td>
<td>23.5 ± 2.0</td>
</tr>
<tr>
<td>MMSE (12 months)</td>
<td>29.1 ± 1.2</td>
<td>26.8 ± 4.3</td>
<td>21.0 ± 4.1</td>
</tr>
<tr>
<td>CDR</td>
<td>0</td>
<td>0.5</td>
<td>0.5 or 1</td>
</tr>
</tbody>
</table>

From table (5.1), it can be seen that patients suffering from AD had the steepest decrease in their cognitive function (MMSE) between the two time instants, followed by MCI patients. Both had an increase in variability, also. Regarding the Control Normal individuals, these maintained their performance.

5.4 Results

5.4.1 Yakushev Normalisation

As referred in chapters 2 and 4, Yakushev Normalisation was one of the first procedures applied in this thesis. Its main effect is the reduction of image artifacts, namely false hypermetabolisation areas, that occur in Alzheimer patients images whenever global mean normalisation is applied. Furthermore, hypometabolic areas in AD patients become more noticeable (see section 4 of chapter 4). The effect of this step was studied in detail through the calculation of average regional voxel intensities before and after Yakushev normalisation, for all the classification tasks (CN vs. AD, CN vs. MCI and MCI vs. AD).

Figures (5.1) shows the outcome of Yakushev Normalisation regarding regional average voxel intensities in the case of CN vs. AD, CN vs. MCI and MCI vs. AD, respectively, at the baseline, obtained with one fold of CV. The ten regions have previously been chosen by an expert, and correspond to areas of the brain that are commonly associated with Alzheimer’s Disease, thus having more pronounced differences between clinical groups. Hence, they are not expected to have the aforementioned hypermetabolic artifacts in the AD group. Impressively, however, the superior anterior cingulate (region 9) presents a higher AD group average than the CN group with CGM normalisation in figure (5.1), even if not statistically significant (see table (5.2)). With this normalisation, only two regions are significantly different between AD and CN groups: the right dorsolateral parietal (r. 8) and the posterior cingulate and precuneus (r. 10). Contrarily, with Yakushev Normalisation, there is a conspicuous effect in all regions, and group differences are all statistically significant, with extremely low p-values (see table (5.1)). In both normalisations, the region with higher difference is the posterior...
cingulate and precuneus.

**Figure 5.1:** Comparison between CGM Normalisation (a) and Yakushev Normalisation (b) in CN vs AD classification. Green dots - CN; Red dots - AD. Comparison between CGM Normalisation (c) and Yakushev Normalisation (d) in CN vs MCI classification. Green dots - CN; Red dots - MCI. Comparison between CGM Normalisation (e) and Yakushev Normalisation (f) in MCI vs AD classification. Green dots - MCI; Red dots - AD. Regional average voxel intensities measured across each group. In (b),(d) and (f) the difference between groups is much more noticeable, as expected. Results are presented for the following regions: 1 - Left lateral temporal; 2 - Right lateral temporal; 3 - Left mesial temporal; 4 - Right mesial temporal; 5 - Inferior frontal gyrus/Orbitofrontal; 6 - Inferior anterior cingulate; 7 - Left dorsolateral parietal; 8 - Right dorsolateral parietal; 9 - Superior anterior cingulate; 10 - Posterior cingulate and precuneus.

Table (5.2) displays p-values obtained with t-tests in every classification modality for the various regions under analysis and for each normalisation, based on a run of 10 folds with cross-validation. In Yakushev Normalisation, each fold is associated to a different reference cluster, which introduces variability. For simplicity, standard deviation is omitted for this case, but it ranged from $10^{-15}$ to $10^{-4}$. 

69
Table 5.2: Regional voxel intensity average differences between clinical groups according to each normalisation procedure. P-values are presented for each brain region: 1 - Left lateral temporal; 2 - Right lateral temporal; 3 - Left mesial temporal; 4 - Right mesial temporal; 5 - Inferior frontal gyrus/Orbitofrontal; 6 - Inferior anterior cingulate; 7 - Left dorsolateral parietal; 8 - Right dorsolateral parietal; 9 - Superior anterior cingulate; 10 - Posterior cingulate and precuneus. Region p-values with statistical significance are in bold

<table>
<thead>
<tr>
<th>Region</th>
<th>CN vs. AD</th>
<th>CN vs. MCI</th>
<th>MCI vs. AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.9 × 10^{-1}</td>
<td>3.8 × 10^{-8}</td>
<td>8.3 × 10^{-1}</td>
</tr>
<tr>
<td>2</td>
<td>1.1 × 10^{-1}</td>
<td>1.7 × 10^{-9}</td>
<td>4.1 × 10^{-1}</td>
</tr>
<tr>
<td>3</td>
<td>1.4 × 10^{-1}</td>
<td>7.9 × 10^{-9}</td>
<td>1.6 × 10^{-1}</td>
</tr>
<tr>
<td>4</td>
<td>2.0 × 10^{-2}</td>
<td>8.4 × 10^{-11}</td>
<td>4.0 × 10^{-2}</td>
</tr>
<tr>
<td>5</td>
<td>9.6 × 10^{-1}</td>
<td>1.1 × 10^{-6}</td>
<td>2.9 × 10^{-1}</td>
</tr>
<tr>
<td>6</td>
<td>4.5 × 10^{-1}</td>
<td>9.8 × 10^{-7}</td>
<td>3.3 × 10^{-3}</td>
</tr>
<tr>
<td>7</td>
<td>4.0 × 10^{-2}</td>
<td>8.9 × 10^{-10}</td>
<td>3.8 × 10^{-1}</td>
</tr>
<tr>
<td>8</td>
<td>3.3 × 10^{-3}</td>
<td>1.0 × 10^{-10}</td>
<td>1.4 × 10^{-1}</td>
</tr>
<tr>
<td>9</td>
<td>3.6 × 10^{-1}</td>
<td>1.9 × 10^{-4}</td>
<td>1.8 × 10^{-1}</td>
</tr>
<tr>
<td>10</td>
<td>2.3 × 10^{-10}</td>
<td>5.4 × 10^{-16}</td>
<td>4.10^{-4}</td>
</tr>
</tbody>
</table>

5.4.2 Feature selection analysis

As mentioned in chapter 4, feature selection consisted of choosing the K best patches according to an absolute t-value sorting. For each classification task (CN vs. AD, CN vs. MCI and MCI vs. AD) the mean t-value of cubes of varying edge dimensions was computed in a voxelwise manner. Figures (5.2), (5.3) and (5.4) present the areas of the brain more likely to be selected as patches when an edge of 3 voxels is used, for the three classification tasks, respectively. As expected, CN and AD groups presented the most marked differences, mainly in the posterior cingulate and precuneus area, but also in the left dorsolateral parietal and right dorsolateral parietal. CN and MCI groups were those whose glucose consumption differences were less noticeable. Table (5.3) complements the images below, carrying information about the highest mean t-value in each classification task, as also the respective brain area. Results are in accordance with literature, showing the posterior cingulate to be the most discriminative area [68].

Furthermore, the percentage of patches that fell in each region was calculated for each combination of patch parameters (edge length (l), distance threshold (d) and number of patches (k) - 128 combinations in total). As the number of patches and distance threshold between them increases, their positions become more widespread. In those cases, many of them have low mean t-value, being less distinctive between patient groups, and thus accounting for the ambiguity in classification.
Figure 5.2: Mean cubic t-value for CN vs. AD group comparison at the baseline. Each voxel intensity depicts the mean t-value of a cube with edge $l = 3$.

Figure 5.3: Mean patch t-value for CN vs. MCI group comparison in the baseline. Each voxel intensity depicts the mean t-value of a cube with edge $l = 3$.

Figure 5.4: Mean patch t-value for MCI vs. AD group comparison at the baseline. Each voxel intensity depicts the mean t-value of a cube with edge $l = 3$.

Table 5.3: Most discriminative regions by maximum patch average t-values, for each classification task.

<table>
<thead>
<tr>
<th></th>
<th>highest mean t-value</th>
<th>region</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>11.8 ±0.8</td>
<td>Posterior cingulate and precuneus</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>5.8 ±0.8</td>
<td>Posterior cingulate and precuneus</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>8.0 ±0.5</td>
<td>Posterior cingulate and precuneus</td>
</tr>
</tbody>
</table>
5.4.3 SVM with best features

For the baseline, the best balanced accuracy was obtained for CN vs. AD (85.02%, 3.41%), with 1000 features. The MCI vs. AD best result was (69.20%, 2.39%) with 5000 features and CN vs. MCI obtained (59.93%, 2.74%), with 1000 features. SVM performance was relatively constant with increasing number of features, as expected, and differences of performance inside each modality (CN vs. AD, CN vs. MCI and MCI vs. AD) are not statistically significant (see figure (5.5)). Remarkably, for numbers as small as 100 features, the results present already a stable behaviour. In agreement with this results, ROC curves for each modality and number of features did not show relevant differences for different number of features (not shown). The differences between each modality are all statistically significant.

With respect to 12 months, it should be noted that there is an apparent improvement for CN vs. AD (89.40%, 2.03%, 5000 features) and CN vs MCI, although to a less extent (61.14%, 3.84%, 500 features). For MCI vs. AD there was an apparent decrease (66.81%, 3.08%, 1000 features), but none of these differences was statistically significant. However, ROC curves show differences between the two time instants (except in MCI vs. AD, see figure (5.6)). What concerns differences between each modality, all except CN vs. MCI and MCI vs. AD (p=0.27) are statistically significant.

These results are in line with the idea that as time progresses, differences between the CN group and the AD group tend to increase, while the MCI group presents oscillatory results. This reflects in a bigger accuracy between the first two for the 12 months and a stabilised one for comparisons between MCI and AD.

![Figure 5.5: Balanced accuracy and standard deviation (%) for each modality in SVM with best features for the baseline (a) and 12 months (b) with number of features ∈ {100, 500, 1000, 5000, 15000}.](image)
Figure 5.6: SVM with best features ROC curves and respective areas under the curve for baseline and 12 months in each classification task: (a) - CN vs. AD; (b) - CN vs. MCI; (c) - MCI vs. AD.

The sensitivity and specificity obtained for the best number of features in each classification modality at the baseline and 12 months, emphasised no deep change between these two instants, even if CN vs. AD and CN vs. MCI presented slightly better results at the 12 months (not shown).

5.4.4 SVM with patches

The best result for SVM with patches at the baseline was (86.74%,2.06%), for CN vs. AD. The difference against the worst result for CN vs. AD is not statistically significant (p=0.19). CN vs. MCI best and worse results were (61.08%,2.19%) and (54.90%,4.31%), respectively, the difference being not statistically significant (p=0.90). Finally, MCI vs. AD obtained a maximum (69.80%,1.38%) and minimum (64.25%,3.56%), again with no statistical significance (p=0.23), as shown in table (5.4). Sensitivity and specificity analysis shows that the latter is better for CN vs. MCI and MCI vs. AD modalities, while for CN vs. AD they are similar (not shown).

For the 12 months, the results seem better for CN vs. AD, but without statistical significance. CN vs. AD obtained (90.14%,3.66%), CN vs. MCI (65.80%,9.05%) and MCI vs. AD (69.76%,7.97%), as displayed in table (5.5). In terms of sensitivity and specificity, the same trend as at the baseline is verified. On the other hand, ROC curve analysis suggests an improvement for the 12 months for CN vs. AD (p=0.01) and a worse result for MCI vs. AD (p=0.05) (see figure (5.7)).
Table 5.4: Best and worst patch parameter combination for SVM with patches at the baseline

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>100</td>
<td>5</td>
<td>3</td>
<td>86.74%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>100</td>
<td>5</td>
<td>5</td>
<td>79.79%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>100</td>
<td>9</td>
<td>5</td>
<td>61.08%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>54.90%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td>69.80%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>1</td>
<td>5</td>
<td>-</td>
<td>64.25%</td>
</tr>
</tbody>
</table>

Table 5.5: Best and worst patch parameter combination for SVM with patches for the 12 months.

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>20</td>
<td>3</td>
<td>10</td>
<td>90.14%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>100</td>
<td>5</td>
<td>5</td>
<td>83.97%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td>65.79%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>56.28%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>50</td>
<td>7</td>
<td>10</td>
<td>69.76%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>50</td>
<td>3</td>
<td>2</td>
<td>61.49%</td>
</tr>
</tbody>
</table>

An aspect that stands out when looking at the tables above is the constancy of results with the varying patch parameters, in both time instants (see tables (5.4) and (5.5)).

![ROC curves and respective areas under the curve for baseline and 12 months in each classification task with the best parameters](image)

Figure 5.7: ROC curves and respective areas under the curve for baseline and 12 months in each classification task with the best parameters: (a) - CN vs. AD; (b) - CN vs. MCI; (c) - MCI vs. AD.

5.4.5 CKNN

CKNN best results for the baseline were (84.83%, 2.70%) for CN vs. AD, (58.46%, 2.30%) for CN vs. MCI and (64.85%, 1.89%) for MCI vs. AD. For the 12 months, these were better, although not statistically significant. Concretely, (90.50%, 3.09%), (60.68%, 2.62%) and (67.37%, 0.60%) were
obtained for CN vs. AD, CN vs. MCI and MCI vs. AD, respectively, as shown in tables (5.6) and (5.7).

Two phenomena happen with CKNN that should be highlighted. Firstly, although the balanced accuracy values seem to be comparable to the previous algorithms based on SVM, sensitivity and specificity results show an unalike behaviour, particularly for CN vs. MCI and MCI vs. AD. The latter even presents a higher specificity than CN vs. AD, in both time instants (see figure (5.8)). Secondly, although not statistically significant, there seems to be an accuracy decrease with distance threshold and patch size (see tables (5.6) and (5.7) and figure (5.9)). This may be related to a reduction in sensitivity (consequently an increase in the false negative rate), which is not shown. In fact, as the distance threshold and patch size increase, more "false positive" patches appear in positive bags. These are in reality closer to negative patches than positive ones, thus increasing the risk of false negatives.

**Table 5.6:** Best and worst patch parameter combination for CKNN at the baseline.

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>84.83%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>20</td>
<td>7</td>
<td>10</td>
<td>75.93%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>10</td>
<td>9</td>
<td>3</td>
<td>58.46%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>20</td>
<td>7</td>
<td>10</td>
<td>52.42%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>64.85%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>50</td>
<td>3</td>
<td>10</td>
<td>55.53%</td>
</tr>
</tbody>
</table>

**Table 5.7:** Best and worst patch parameter combination for CKNN for the 12 months.

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>1</td>
<td>3</td>
<td>-</td>
<td>90.50%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>50</td>
<td>7</td>
<td>10</td>
<td>82.10%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>20</td>
<td>9</td>
<td>3</td>
<td>60.68%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>20</td>
<td>9</td>
<td>10</td>
<td>53.90%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>1</td>
<td>5</td>
<td>-</td>
<td>67.37%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>50</td>
<td>9</td>
<td>10</td>
<td>57.13%</td>
</tr>
</tbody>
</table>

**Figure 5.8:** Sensitivity and specificity for the best parameter combination in each modality at the baseline (a) and 12 months (b).
5.4.6 Diverse Density

Diverse Density best results were obtained for CN vs. AD for both time instants: (85.30%,0.82%) and (89.91%,1.71%), respectively. CN vs. MCI reached (59.60%,-) and (64.21%,-), respectively, and MCI vs. AD attained (66.21%,1.52%) and (66.21%,1.52%). It should be noted that due to the long running time of this algorithm, sometimes it was impossible to perform more than one run of CV, which ruled out the computation of standard deviation.

The performance of this classifier seems not to be affected by the patch parameters. Furthermore, the differences between the two time measurements are not statistically significant, even if there is an apparent improvement for the 12 months in the CN vs. MCI and CN vs. AD cases. Sensitivity and specificity of the best patch parameters are depicted in figure (5.10).

Table 5.8: Best and worst patch parameter combination for DD at the baseline.

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>85.30%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>76.82%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>2</td>
<td>9</td>
<td>3</td>
<td>59.60%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>49.84%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>71.14%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>5</td>
<td>9</td>
<td>2</td>
<td>63.38%</td>
</tr>
</tbody>
</table>

Table 5.9: Best and worst patch parameter combination for DD at 12 months.

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>1</td>
<td>5</td>
<td>-</td>
<td>89.91%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>2</td>
<td>9</td>
<td>10</td>
<td>82.10%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>64.21%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>56.48%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>66.21%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>1</td>
<td>7</td>
<td>-</td>
<td>63.78%</td>
</tr>
</tbody>
</table>
MILES results are shown in tables (5.10) and (5.11). CN vs. AD was once again the best case at the baseline (85.90%,3.38%) and 12 months (90.00%,2.52%), followed by MCI vs. AD (68.68%,1.57% and 69.04%,1.83%, respectively) and CN vs. MCI (61.83%,1.13% and 63.82%,1.32%, respectively).

**Table 5.10:** Best and worst patch parameter combination for MILES at the baseline.

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>85.90%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>100</td>
<td>5</td>
<td>10</td>
<td>78.51%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>61.83%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>1</td>
<td>3</td>
<td>-</td>
<td>53.68%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>10</td>
<td>3</td>
<td>10</td>
<td>68.68%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>20</td>
<td>9</td>
<td>3</td>
<td>61.77%</td>
</tr>
</tbody>
</table>

**Table 5.11:** Best and worst patch parameter combination for MILES for the 12 months.

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>90.00%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>10</td>
<td>9</td>
<td>5</td>
<td>83.41%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>20</td>
<td>7</td>
<td>3</td>
<td>63.82%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>50</td>
<td>9</td>
<td>10</td>
<td>55.99%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>50</td>
<td>9</td>
<td>5</td>
<td>69.04%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>20</td>
<td>5</td>
<td>3</td>
<td>57.45%</td>
</tr>
</tbody>
</table>

An interesting aspect of MILES was its ability to correctly classify training instances. Figure (5.11) shows the estimated probability distribution of weights attributed to training instances in positive and negative bags. Weights with positive value are considered positive for AD, and negative otherwise. It can be seen that instances belonging to negative bags are primarily classified as negative, thus having few false positives. As a matter of fact, at the baseline and for bags with 2 patches, there were on average 33.49 ± 4.31% false positives for CN vs. AD at the baseline, while CN vs. MCI had an average 34.05±5.07% and MCI vs. AD 30.15±1.59%. At the 12 months, the percentage dropped significantly for CN vs. AD, showing 19.27 ± 4.41% false positives, while for CN vs. MCI showed 27.74 ± 6.91%
and MCI vs. AD percentages got worse, with 34.63 ± 1.81%. The vast majority of these false positive training instances seems to be concentrated around zero (not shown), which leads to suppose they would be classified as void weights if SVM norm 1 was used instead. With respect to positive bags, it is not possible to infer the percentage of false negative instances nor false positive ones, since MIL allows bags to have negative instances.

Figure 5.11: Example of estimated probability distribution of instance weights belonging to negative bags (red) and positive bags (green), obtained with one fold of CV for CN vs. AD at the baseline. Each bag had 2 patches and weights were scaled with the difference between the max and min weights.

Figure (5.12) shows the ROC curves and sensitivity/specificity obtained for each modality in both time instants, respectively. ROC analysis shows that CN vs. MCI improved at the 12 months (p=7.3×10⁻⁵).
Figure 5.12: ROC curves and respective areas under the curve for the baseline and 12 months in each classification task (MILES): (a) - CN vs. AD; (b) - CN vs. MCI; (c) - MCI vs. AD.

5.4.8 YARDS

YARDS best results were, once again, obtained for the CN vs. AD case, with (86.05%,2.41%) for the baseline and (89.23%,1.95%) for the 12 months. CN vs. MCI best result was (63.15%,1.39%) and (66.85%,4.13%), for the baseline and 12 months, respectively. Finally, MCI vs. AD reached (69.99%,1.81%) and (70.21%,1.15%) for the two time instants, respectively. The results seem to be stable for the various patch parameters and the differences between the two time instants are not statistically significant, although in the CN vs. AD and CN vs. MCI modalities there is an apparent improvement in accuracy. ROC analysis also suggests this trend, showing that temporal differences are significant for CN vs. AD (p=1×10^{-10}) and CN vs. MCI little above the significance level (p=0.07). What concerns sensitivity and specificity, both present similar values at the baseline and 12 months. It should be noted that YARDS also presented the possibility of training instance classification, as MILES, with similar values and trends (not shown).

Table 5.12: Best and worst patch parameter combination for YARDS at the baseline.

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>86.05%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>50</td>
<td>5</td>
<td>10</td>
<td>75.87%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>20</td>
<td>9</td>
<td>3</td>
<td>63.15%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>2</td>
<td>9</td>
<td>10</td>
<td>51.92%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>15</td>
<td>7</td>
<td>10</td>
<td>69.99%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>10</td>
<td>9</td>
<td>5</td>
<td>60.76%</td>
</tr>
</tbody>
</table>
### Table 5.13: Best and worst patch parameter combination for YARDS at 12 months.

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>1</td>
<td>3</td>
<td>-</td>
<td>89.23%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>50</td>
<td>9</td>
<td>3</td>
<td>83.48%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>50</td>
<td>9</td>
<td>10</td>
<td>66.85%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>55.38%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>15</td>
<td>7</td>
<td>3</td>
<td>70.21%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>50</td>
<td>9</td>
<td>10</td>
<td>61.36%</td>
</tr>
</tbody>
</table>

![ROC curves and respective areas under the curve for the baseline and 12 months in each classification task (YARDS): (a) - CN vs. AD; (b) - CN vs. MCI; (c) - MCI vs. AD.](image)

### 5.4.9 BARTMIP

BARTMIP results show a similar pattern to the ones presented before and are displayed in table (5.15). As it can be seen in figure (5.14), the number of clustered groups does not change the performance of the algorithm, which is in accordance with results from the literature on other data sets, such as Musk1 and Musk2 [132]. Besides, results do not seem to depend on the patch parameters, even if the best results are very often obtained for a small number of patches K.

The effect of increasing the number of centroids on average cluster purity and entropy was also analysed, and is in agreement with the results of [132]. In fact, there is an increase in purity with the number of centroids and a decrease in entropy, with statistical significance. Remarkably, with only two clusters, the average purity for CN vs. AD is in the order of 90%, which indicates that the unsupervised step can uncover much of the structure of the data set. Increasing the number of medoids, apparently, does not reflect in the accuracy of the classifier (see figure (5.14)). Overall, however, there is an agreement between purity and accuracy results, that is, classification between clinical groups with higher purity presents also higher accuracy. On the other hand, entropy decreases with the number of
centroids (see figure (5.15)).

Table 5.14: Best and worst patch parameter combination for BARTMIP with patches at the baseline.

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th># centroids</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>50</td>
<td>87.10%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>20</td>
<td>7</td>
<td>10</td>
<td>30</td>
<td>77.11%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>50</td>
<td>5</td>
<td>5</td>
<td>50</td>
<td>63.83%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>15</td>
<td>7</td>
<td>5</td>
<td>50</td>
<td>51.97%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>50</td>
<td>9</td>
<td>10</td>
<td>30</td>
<td>71.32%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td>50</td>
<td>59.72%</td>
</tr>
</tbody>
</table>

Table 5.15: Best and worst patch parameter combination for BARTMIP with patches at 12 months.

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th># centroids</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>1</td>
<td>3</td>
<td>-</td>
<td>10</td>
<td>90.98%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>50</td>
<td>9</td>
<td>5</td>
<td>50</td>
<td>82.28%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>30</td>
<td>65.81%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>30</td>
<td>53.61%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>20</td>
<td>3</td>
<td>3</td>
<td>10</td>
<td>70.19%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>15</td>
<td>7</td>
<td>3</td>
<td>30</td>
<td>60.06%</td>
</tr>
</tbody>
</table>

Figure 5.14: Balanced accuracy and standard deviation (%) for different number of centroids with BARTMIP at the baseline (a) and 12 months (b).
Figure 5.15: Average purity and entropy variation with centroid number for CN vs. AD, CN vs. MCI and MCI vs. AD at the baseline. Results refer to average of 10 runs of 10 fold CV, using 2 patches per patient (size 3x3x3 and distance threshold of 10).

ROC analysis in figure (5.16) shows that there is a significant increase (p=0.02) in the AUC for CN vs. AD between the baseline and the 12 months, while for CN vs. MCI and MCI vs. AD, differences are not significant.

Figure 5.16: ROC curves and respective areas under the curve for the baseline and 12 months in each classification task with the best parameters: (a) - CN vs. AD; (b) - CN vs. MCI; (c) - MCI vs. AD.

5.4.10 Longitudinal Analysis

Longitudinal results are very similar to the cross-sectional ones at the 12 months. This seems to reflect that adding temporal information improves relatively to the baseline, but not relatively to the 12 months. Regarding the two algorithms tested for this analysis, accuracy differences between SVM
with patches and MILES are not statistically significant. On the contrary, ROC curves for CN vs. AD and CN vs. MCI reveal a better performance of SVM with patches (see figure (5.17)).

**Table 5.16:** Best and worst patch parameter combination for SVM with patches (longitudinal analysis)

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>20</td>
<td>3</td>
<td>10</td>
<td>91.21%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>2</td>
<td>9</td>
<td>10</td>
<td>85.30%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>64.85%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>10</td>
<td>7</td>
<td>5</td>
<td>54.84%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>69.39%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>50</td>
<td>3</td>
<td>10</td>
<td>64.01%</td>
</tr>
</tbody>
</table>

**Table 5.17:** Best and worst patch parameter combination for MILES (longitudinal analysis)

<table>
<thead>
<tr>
<th>Modality</th>
<th>K</th>
<th>l</th>
<th>d</th>
<th>b.acc</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. AD</td>
<td>best</td>
<td>1</td>
<td>5</td>
<td>-</td>
<td>90.10%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>20</td>
<td>7</td>
<td>10</td>
<td>84.16%</td>
</tr>
<tr>
<td>CN vs. MCI</td>
<td>best</td>
<td>20</td>
<td>7</td>
<td>5</td>
<td>63.77%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>50</td>
<td>9</td>
<td>5</td>
<td>53.85%</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>best</td>
<td>20</td>
<td>3</td>
<td>2</td>
<td>68.15%</td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td>15</td>
<td>9</td>
<td>10</td>
<td>58.29%</td>
</tr>
</tbody>
</table>

**Figure 5.17:** ROC curves and respective areas under the curve for the longitudinal analysis: (a) - CN vs. AD; (b) - CN vs. MCI; (c) - MCI vs. AD.

5.4.11 Comparison of Classifiers

ROC curve analysis shows that no significant difference exists among SVM, MILES, YARDS and BARTMIP at the baseline. From those, SVM with the best features presents the highest area under the curve (AUC=0.92), and MILES the smallest one (AUC=0.90). At the 12 months, the best ROC curve
is still that of SVM with the best features (AUC=0.96), followed by SVM with patches (AUC=0.95).
The worst performance belongs, again, to MILES (AUC=0.89). The difference between SVM with
the best features against all the other algorithms (except SVM with patches) is significant (YARDS
p=0.003; MILES p=0.002; BARTMIP p=0.027). One can also see that ROC curves at the 12 months
are in general better than at the baseline. Results are displayed in figure (5.18).

![ROC comparison for SVM with best features, SVM with patches, MILES, YARDS and BARTMIP
at the baseline (a) and 12 months (b) for CN vs. AD.](image)

**Figure 5.18**

As to CN vs. MCI, SVM with patches has the best ROC at the baseline (AUC=0.67), a value
which is significantly better than all the others (SVM with best features p = 0.02; YARDS p=0.02;
MILES p=1.3×10^{-3}; BARTMIP p=0.01). SVM with best features is also significantly better than
YARDS (p=0.04) and MILES (p=3.7×10^{-3}). At the 12 months, ROC curves seem to behave all in a
similar way, with no significant differences, as depicted in figure (5.19).

![ROC comparison for SVM with best features, SVM with patches, MILES, YARDS and BARTMIP
at the baseline (a) and 12 months (b) for CN vs. MCI.](image)

**Figure 5.19**

Finally, in what concerns MCI vs. AD, SVM with patches is the classifier with the highest AUC
(AUC=0.77) at the baseline, while YARDS (AUC=0.72) and MILES (AUC=0.71) have the worst
values. The differences between the best classifier and the two worst are all significative (all p-values
smaller then 0.03). For the 12 months, SVM with the best features is the classifier with the highest
AUC (0.74), which is only significantly better than YARDS (p=0.02) (see figure (5.20)).
Tables (5.18), (5.19) and (5.20) show the best results of each modality, respectively, for every classifier at the baseline and 12 months, and for SVM with patches and MILES in the longitudinal analysis. One of the first aspects that stands out is the improvement that occurs for every classifier at the 12 months, in CN vs. AD and CN vs. MCI, while MCI vs. AD seems to have a similar performance at the two time instants. Secondly, CKNN presents the worst results for CN vs. AD at the baseline, CN vs. MCI at the baseline and 12 months, and MCI vs. AD at the baseline, even if not statistically significant. Furthermore, the best balanced accuracy attained belongs to SVM with patches in the longitudinal analysis (91.21%) in CN vs. AD. Nonetheless, longitudinal analysis did not show any significant improvement.

Table 5.18: Best results for each classifier (CN vs. AD)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Baseline</th>
<th>12 months</th>
<th>Longitudinal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b.acc</td>
<td>std</td>
<td>b.acc</td>
</tr>
<tr>
<td>SVM with best features</td>
<td>85.02%</td>
<td>3.41%</td>
<td>89.40%</td>
</tr>
<tr>
<td>SVM with patches</td>
<td>86.74%</td>
<td>2.06%</td>
<td>90.14%</td>
</tr>
<tr>
<td>CKNN</td>
<td>84.83%</td>
<td>2.70%</td>
<td>90.50%</td>
</tr>
<tr>
<td>DD</td>
<td>85.30%</td>
<td>0.82%</td>
<td>89.91%</td>
</tr>
<tr>
<td>MILES</td>
<td>85.90%</td>
<td>3.38%</td>
<td>90.00%</td>
</tr>
<tr>
<td>YARDS</td>
<td>86.05%</td>
<td>2.41%</td>
<td>89.23%</td>
</tr>
<tr>
<td>BARTMIP</td>
<td>87.10%</td>
<td>1.83%</td>
<td>90.98%</td>
</tr>
</tbody>
</table>

Table 5.19: Best results for each classifier (CN vs. MCI)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Baseline</th>
<th>12 months</th>
<th>Longitudinal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b.acc</td>
<td>std</td>
<td>b.acc</td>
</tr>
<tr>
<td>SVM with best features</td>
<td>59.93%</td>
<td>2.74%</td>
<td>61.14%</td>
</tr>
<tr>
<td>SVM with patches</td>
<td>61.08%</td>
<td>2.19%</td>
<td>65.79%</td>
</tr>
<tr>
<td>CKNN</td>
<td>58.46%</td>
<td>2.30%</td>
<td>60.68%</td>
</tr>
<tr>
<td>DD</td>
<td>59.60%</td>
<td>-%</td>
<td>64.21%</td>
</tr>
<tr>
<td>MILES</td>
<td>61.83%</td>
<td>1.13%</td>
<td>63.82%</td>
</tr>
<tr>
<td>YARDS</td>
<td>63.15%</td>
<td>1.39%</td>
<td>66.85%</td>
</tr>
<tr>
<td>BARTMIP</td>
<td>63.83%</td>
<td>0.03%</td>
<td>65.81%</td>
</tr>
</tbody>
</table>
Table 5.20: Best results for each classifier (MCI vs. AD)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Baseline b.acc</th>
<th>Baseline std</th>
<th>12 months b.acc</th>
<th>12 months std</th>
<th>Longitudinal b.acc</th>
<th>Longitudinal std</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM with best features</td>
<td>69.20%</td>
<td>2.39%</td>
<td>66.81%</td>
<td>3.08%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SVM with patches</td>
<td>69.80%</td>
<td>1.38%</td>
<td>69.76%</td>
<td>0.64%</td>
<td>69.39%</td>
<td>2.37%</td>
</tr>
<tr>
<td>CKNN</td>
<td>64.85%</td>
<td>1.89%</td>
<td>67.37%</td>
<td>3.40%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DD</td>
<td>71.14%</td>
<td>1.27%</td>
<td>66.21%</td>
<td>1.52%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MILES</td>
<td>68.68%</td>
<td>1.57%</td>
<td>69.04%</td>
<td>1.83%</td>
<td>68.15%</td>
<td>3.50%</td>
</tr>
<tr>
<td>YARDS</td>
<td>69.99%</td>
<td>1.81%</td>
<td>70.21%</td>
<td>5.21%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BARTMIP</td>
<td>71.32%</td>
<td>1.25%</td>
<td>70.19%</td>
<td>1.73%</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

5.4.12 Simulation of non-registered images

Results for the simulation of non-registered images show that MILES can have competitive results for the CN vs. AD case, even if worse than the alternative approach taken in this thesis. In particular, amongst the various tests performed, the one with best results was obtained with 1 random discriminative patch (the only one) per patient in both training and test sets (81.12%, 3.21%). Some general trends can also be inferred. First, when the test patients do not have patches extracted from discriminative areas, the results are far worse, namely the specificity. This effect is shown in figure (5.21) (1,10,1, 1,10,0,10 and 5,10,0,10).

![Figure 5.21](image)

Figure 5.21: The decrease in specificity when no discriminative patches are extracted in test patients. Balanced accuracy, sensitivity and specificity obtained for various tests. Each test is defined by four numbers x, y, z, w, where x is the number of discriminative patches per patient in the training set, y is the total number of patches per patient in the training set, z is the number of discriminative patches per patient in the test set and w the total number of patches per patient in the test set. Results were obtained with 10 runs of 10 fold cross-validation.

This behaviour is in fact part of a more general tendency, which is the increase of specificity with the number of aleatory discriminative patches in test patients. This happens at the expense of a little reduction of sensitivity (see figure (5.22)).
Figure 5.22: The increase of specificity with the number of discriminative patches in test patients. Balanced accuracy, sensitivity and specificity obtained for various tests. Each test is defined by four numbers $x, y, z, w$, where $x$ is the number of discriminative patches per patient in the training set, $y$ is the total number of patches per patient in the training set, $z$ is the number of discriminative patches per patient in the test set and $w$ the total number of patches per patient in the test set. Results were obtained with 10 runs of 10 fold cross-validation.

On the other hand, the increase of discriminative patches in training patients increases the sensitivity, at the expense of some decrease in the specificity, as depicted in figure (5.23).

Figure 5.23: The increase of sensitivity with the number of discriminative patches in training patients. Balanced accuracy, sensitivity and specificity obtained for various tests. Each test is defined by four numbers $x, y, z, w$, where $x$ is the number of discriminative patches per patient in the training set, $y$ is the total number of patches per patient in the training set, $z$ is the number of discriminative patches per patient in the test set and $w$ the total number of patches per patient in the test set. Results were obtained with 10 runs of 10 fold cross-validation.
Conclusions and Future Work
The current thesis exploited the use of Multiple Instance Learning, a new machine learning approach, in the automatic diagnosis of AD. CAD of AD is a hot topic for the scientific community, both for its impact in an ever ageing society and for its challenging aspects. As a starting point, it was hypothesised that this paradigm could improve diagnosis accuracy, since it introduces an ambiguity in the process of classification. In fact, people suffering from Alzheimer show brain atrophy, metabolism reduction and other artifacts only in some areas of the brain, which means that "false negative" areas do exist. MIL framework seems to tackle well this ambiguity aspect.

This work constitutes a breakthrough in the field, since the vast majority of research has been developed in the context of supervised learning. Particularly, three main domains of MIL - the instance space based, the bag space based and the embedded space based - in a total of 5 classifiers, were tested and compared to two simple SVM approaches for ADNI data at the baseline and 12 months.

In order to accomplish this approach, several voxel intensity patches were extracted from FDG-PET images. For every algorithm, except one - Diverse Density -, these were directly fed to the classifier. For DD, patch intensity histograms were derived. The patch selection procedure was based on group driven voxelwise t-tests, in descent order of statistic value, thus maximising the potential discrimination between clinical groups (CN, MCI and AD). Several combinations of patch parameters (number, distance threshold and edge size) were tested. Importantly, an enhancing preprocessing step, known as Yakushev Normalisation, was applied to patient images, revealing improved discrimination accuracy, particularly for CN vs. AD and MCI vs. AD.

Simultaneously, one less obvious but very important aspect of MIL was explored, which was its applicability for non-registered images. Indeed, it soon became evident that if this paradigm allows for the use of different patches in different patients, the use of non-registered images is just a step beyond, even if ambitious. Given the short time frame, it was decided not to unregister the (already registered) images, but rather simulate results with different patches per patient and also with randomness included.

Regarding the results, while not outperforming their supervised counterparts, these were competitive with the literature. CN vs. AD was the modality with the best balanced accuracy, as expected. MCI vs. AD systematically outperformed CN vs. MCI results, which indicates that patients in the MCI stage are "closer" to the normal ones. Nonetheless, both modalities show that a stricter definition of MCI stage must be established, so that an early diagnosis becomes possible. One plausible way to accomplish this is the introduction of a distinction between conversor and non-conversor MCI patients, which is indeed being assessed by some authors [65, 155]. Furthermore, it should be noted that all methods seem to be robust against patch parameters change, at least under this selection procedure, which is in disagreement with results obtained by Tong [65, 98]. A possible explanation is the different data type being analysed (PET in this thesis, MRI in Tong work), which might be less prone to redundant information carried by patches. Notably, all classifiers obtained competitive results with only one patch. Even if from an anatomical point of view this seems difficult to grasp, due to patch minute dimensions, this acknowledges the fact that the most discriminative patch is probably the most important. Further research should be done under other conditions, such as randomness,
different number of patches per patient and different location of patches per patient. Interestingly, results with DD, despite using histograms as instances, and CKNN, with no use of SVM, obtained very similar results to the remaining three MIL classifiers and the supervised ones. This might be a hint that improvements in classification, in the future, will probably be achieved with new feature extraction and selection procedures, not with new classifiers. In fact, most of recent research has been centered in the combination of multiple biomarkers, which together may provide more information than on their own. For instance, blood biomarkers, together with MRI or PET, might be used. In addition, one should not discard the use of longitudinal data, which might complement the use of various biomarkers. Although not showing very promising results, and sometimes contradictory ones, it is indeed a field to be further explored.

Regarding ROC curves, these revealed in general an improvement at the 12 months for CN vs. AD and CN vs. MCI, while for MCI vs. AD these were less evident. Besides, supervised SVM with best features and patches obtained often the best Area Under the Curve, while YARDS and MILES showed the worse ones, a fact that did not materialise in a better or worse accuracy, respectively.

In short, comparison between the 5 MIL algorithms showed no significant differences. The only major difference that was identified was a distinct behaviour of CKNN with respect to sensitivity and specificity, which were particularly marked for the CN vs. MCI and MCI vs. AD comparisons. A somehow unexpected fact was the apparent independence of the classifiers regarding the patch parameters, except maybe for CKNN. One should explore other patch selection approaches to check if this homogeneous behaviour is maintained.

In addition, BARTMIP results set forth the use of clustering methods in the context of MIL, namely applied to AD diagnosis. The impact the cluster quality (purity and entropy, among other possible measures) has on accuracy should be further looked into. As a matter of fact, purity and entropy variation with the number of centroids did not show any tangible effect in accuracy inside each classification task. Nonetheless, the parity between accuracy and each classification task accuracy is undeniable.

Diverse Density, even if with competitive results, is not viable due to its fastidious classification process, which requires several weeks for even few patches. EM-DD should demonstrate to be more efficient.

MILES and YARDS, both simple in their theoretical background, have the additional appeal that they are able to label training instances, even if with some false positives and (possibly) false negatives. This aspect should be further investigated and classifiers other than SVM should be tested in the context of AD. Also, the possibility of instance classification in test bags should be analysed in the future.

Concerning the simulation of non-registered images, results should be interpreted with caution, but there seems to be some applicability, even if with less accuracy than with the main approach of this thesis. This, however, is natural, since randomness was introduced, rather than a t statistic driven selection. The number of patches chosen at random from discriminant regions seems to play an important role. Future work should try new combinations of patch parameters and test this approach.
in non registered images. If successful, this would be a great achievement, since registration is a fastidious process. Also, it would be interesting to test a medical driven approach with registered images and aleatory patches extracted from patients under some restrictions given by the doctor.
Bibliography


