TÉCNICO LISBOA

Computer aided diagnosis of Alzheimer's disease from brain images: a method robust to registration errors

Marta Sofia Aranha da Conceição

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor(s): Professor Maria Margarida Campos da Silveira

Examination Committee

Chairperson: Professor Patrícia Margarida Piedade Figueiredo
Supervisor: Professor Maria Margarida Campos da Silveira
Member of the Committee: Professor Ana Luísa Nobre Fred

May 2019

Preface

The work presented in this thesis was performed at the Institute for Systems and Robotics of Instituto Superior Técnico (Lisbon, Portugal), during the period March 2018 - May 2019, under the supervision of Professor Margarida Silveira.

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to Professor Margarida Silveira, for all the useful remarks, suggestions and guidance through multiple and enriching meetings, that contributed immensely to the development of this work.

I would also like to thank, from the bottom of my heart, my friends and family who are always there for me, through ups and downs. To my parents, for constantly doing everything they could so I wouldn't miss out on anything I truly enjoyed, while still making sure that I would always stay focused, work hard and try my best to achieve my goals. To my brother, Vasco, for being the best one I could ever ask for; for always believing in me way more than I do and for all the great memories we've created so far and will continue to form in the future. To my cousins, Rita and João Pedro, for the friendship and patience, for always being available to help me with anything I need, for the awesome childhood we spent together and for making these last few years living under the same roof so much fun and some of the best ones of my life. To my cousin Manel, for being my worst influence in the best possible way and for always keeping our family spirit well alive. To my grandparents, for all the affection, love, concern, courage and fighting spirit. To my uncle and aunt, for constantly pushing me to get out of my comfort zone and for granting me the opportunity to have this amazing life while studying in Lisbon, I truly appreciate it.

To my friends, Ana Sofia Borges, Carolina Carreira, Cláudia Diniz, Constança Durão, Filipa Leal, Joana Freitas, Joana Pinto, João Delgado, João Ramiro, Margarida Silva, Mário Macedo, Patrícia Silva, Rafaela Saraiva, Rita Oliveira, Shin Fujii, Takeshi Ii, Teresa Bucho and YongJun Cho. Whether we've met in our early childhood or the last couple of years, in school or tennis practice, Santarém or Lisbon, Portugal or Denmark, if it wasn't for all the great times we've spent together and the unforgettable memories, for your support, encouragement, and for always being there for me when I needed you, I certainly wouldn't be where I am today as a person nor in terms of my academic journey. Neither of us can tell what the future holds, but you know what they say: while everything else may change, if you choose right, your people will stay the same. Thank you all so much.

Resumo

A doença de Alzheimer (AD) é uma doença neurodegenerativa crónica progressiva que afeta milhões de pessoas mundialmente, maioritariamente idosas. Apesar de não existir cura para a AD, a sua deteção precoce é crucial, pois a sua gestão eficaz pode prevenir a progressão para estadios mais severos. A incerteza inerente ao diagnóstico clínico da AD tem levado à pesquisa de biomarcadores onde a neuroimagiologia, e nomeadamente a tomografia por emissão de positrões (PET), assume um papel central. Para que as imagens adquiridas possam ser utilizadas no diagnóstico assistido por computador da AD, contudo, é geralmente necessário realizar o seu registo para um referencial padrão de coordenadas espaciais. Este processo pode ser complexo, dado que vários desafios, incluindo a variabilidade anatómica inter-sujeito, são encontrados, e possíveis erros de classificação podem advir de uma transformação de coordenadas incorreta. Neste trabalho, para encontrar um método robusto a estes erros de registo, e partindo da abordagem baseada em *textons*, vários métodos foram considerados e aplicados em conjuntos de dados registados ou não. Diferentes *features* foram ainda consideradas, nomeadamente aprendidas usando auto-codificadores empilhados (SSAE) e a intensidade dos voxels, quer extraídos de todo o cérebro, de patches ou de regiões de interesse. A classificação binária entre sujeitos cognitivamente normais, diagnosticados com AD ou com défice cognitivo ligeiro (MCI) foi efetuada de forma dicotómica e foram retiradas conclusões relativas à precisão e robustez dos diferentes métodos, confirmando particularmente a robustez da abordagem baseada em *textons* aplicada às imagens totais do cérebro e das *features* aprendidas usando SSAE.

Palavras-chave: Doença de Alzheimer; Diagnóstico assistido por computador; Tomografia por emissão de positrões; Registo; *Textons*; Classificação

Abstract

Alzheimer's disease (AD) is a chronic progressive neurodegenerative disease affecting millions of people worldwide and prominently the elderly. While there is still no cure for AD, its early detection is crucial, as an effective management of the disease may help prevent the progression to more severe stages. The inherent uncertainty in the clinical diagnosis of AD has driven a search for biomarkers, where brain imaging, and namely positron emission tomography (PET), assumes a key role. For the acquired neuroimaging data to be used for computer aidded diagnosis of AD, however, it's usually necessary to perform image registration to a standard spatial coordinate system. This can be troublesome, as many challenges, including inter-subject anatomical variability, are encountered, so that possible misclassification errors might result from a poor coordinate transformation. In this work, in the attempt to find a method robust to such registration errors, and building from the texton-based approach, several methods were considered and applied on both registered and non-registered datastets. Additional feature representations were considered, namely learned using a stacked sparse autoencoder (SSAE) and the raw voxel intensity values, either extracted from the whole brain, patches or identified regions of interest. Binary classification among cognitively normal subjects, patients diagnosed with AD and with mild cognitive impairment (MCI) was performed in a dichotomous fashion and conclusions regarding the accuracy and robustness of the diferent methods were drawn, particularly confirming the robustness of the texton-based approach applied on the whole brain images and of the learned feature representations using SSAE.

Keywords: Alzheimer's disease; Computer aided diagnosis; Positron emission tomography; Registration; Textons; Classification

Contents

List of Tables

List of Figures

List of Acronyms

1D One dimensional

2D Two dimensional

3D Three dimensional

AChE Acetylcholinesterase

AC Anterior comissure

ADAS Alzheimer's Disease Assessment Scale

ADNI Alzheimer's Disease Neuroimaging Initiative

AD Alzheimer's disease

AE Autoencoder

aMCI Amnestic mild cognitive impairment

ANT Advanced Normalization Tools

APOE Apolipoprotein E

APP Amyloid precursor protein

ASL Arterial spin labeling

AVLT Auditory Verbal Learning Test

Aβ Beta-amyloid

BFGS Broyden-Fletcher-Goldfarb-Shanno

c-MCI Converter mild cognitive impairment

CAD Computer aided diagnosis

CBF Cerebral blood flow

CDR Clinical Dementia Rating

CDT Clock Drawing Test

CERAD Consortium to Establish a Registry for Alzheimer's Disease

CNN Convolutional neural network

CN Cognitively normal

CSF Cerebrospinal fluid

DBM Deep Boltzmann machine

DLB Dementia with Lewy Bodies

DS Down's syndrome

DTI Diffusion tensor imaging

EC Entorhynal cortex

ELM Extreme learning machines

ET Extremely randomized trees

FAD Familial Alzheimer's disease

FAQ Functional Activities Questionnaire

FA Fractional anisotropy

FDG Fluorodeoxyglucose

FDR Fisher Discriminant Ratio

FTD Frontotemporal dementia

GDA Gaussian discriminant analysis

GDS Global Dementia Scale

GHI Generalized histogram intersection

GL-CHF Gauss-Laguerre circular harmonic functions

GM Gray matter

HAMMER Hierarchical attribute matching mechanism for elastic registration

HC Hippocampus

ICAm Independent component analysis on means

ICBM International Consortium of Brain Mapping

ICV Inter-class variance

IVM Import vector machines

k-NN k-Nearest neighbors

KKT Karush-Kuhn-Tucker

KL Kullback-Leibler

LDA Linear discriminant analysis

LEP Local energy patterns

LM Leung and Malik

LM Logic Memory

LoG Laplacian of Gaussian

MCI Mild cognitive impairment

MD Mean diffusivity

MICNN Multi-instance convolutional neural network

MI Mutual information

MKL Multiple kernel learning

mMCI Multimodal mild cognitive impairment

MMSE Mini-Mental State Examination

MNI Montreal Neurological Institute

MRI Magnetic resonance imaging

mRMR Minimum redundancy maximum relevance

MR Magnetic resonance

NGF Non-linear graph fusion

NTB Neuropsychological Test Battery

PCA Principal component analysis

PC Posterior comissure

PET Positron emission tomography

PIB Pittsburgh compound B

PLS Partial least squares

PS1 Presenilin 1

PS2 Presenilin 2

RBF Radial basis function

RBM Restricted Boltzmann machine

rCBF Regional cerebral blood flow

rCMRglc Regional cerebral glucose metabolism

RELM Regularized extreme learning machines

RF Random forests

ROC Receiver operating characteristic

ROI Region of interest

SAD Sporadic Alzheimer's disease

SAE Sparse autoencoder

SCG Scaled conjugate gradient

sMCI Stable mild cognitive impairment

sMRI Structural magnetic resonance imaging

SNIPE Scoring by non-local image patch estimator

SPECT Single-photon emission computed tomography

SPF Separation power factor

SPM Statistical parametric mapping

SRC Sparse representation-based classifier

SSAE Stacked sparse autoencoder

SVM Support vector machines

VBM Voxel-based morphometry

VI Voxel intensity

WM White matter

ZD Zernicke descriptors

Chapter 1

Introduction

1.1 Motivation

Alzheimer's disease (AD) is an ultimately fatal neurodegenerative disease, affecting millions of people worldwide and prominently the elderly [1]. As a form of dementia, the disease associated symptoms include difficulties with memory, language, problem-solving and other cognitive skills that affect the ability to perform everyday activities [1],[2]. According to a 2018 study by the Alzheimer's Association [3], in particular regarding the USA and based on a census from 2010, an estimated 5.7 million of Americans of all ages have AD (which includes an estimated 5.5 million people age 65 and older and 3.4 million women) and by 2050 this prevalence is projected to grow to 13.8 million [1],[4]. Another study states that in 2017 approximately 6.08 million Americans had either clinical AD or its early prodromal stage designated as mild cognitive impairment (MCI) and forecasts that this number will grow to 15.0 million by 2060 [5],[6],[7]. This denomination of MCI corresponds to a stage where the exhibited cognitive decline is greater than expected due to normal aging, but where everyday activities can still be performed without notable impairment, so that the criteria for dementia is not fulfilled. The same study also reports that in 2017 around 46.7 million people had preclinical AD, although many may not progress to clinical disease during their lifetimes [5]. This designation of preclinical AD refers to an earlier stage whose existence is still speculative, characterized by the disease associated brain changes, namely amyloidosis, neurodegeneration, or both [5].

Although there is still no cure for Alzheimer's, and even though the genetic component also plays a role in its etiopathogenesis, the fact that many of its risk factors are modifiable and that some treatments available can help prevent more severe stages of the disease to be entered, makes it extremely important to be able to detect the disease still at its early stages [1]. While the diagnosis of MCI still faces some challenges, neuroimaging tools can not only corroborate it but also aid in predicting which such cases will convert to AD [8]. Furthermore, these tools prove to be useful in establishing a differential diagnosis of AD against other types of dementia and again can corroborate its diagnosis in symptomatic patients if the disease's characteristic metabolic impairment pattern, cerebral atrophy or specific protein deposition is observed through different neuroimaging modalities [9]. Computer aided diagnosis (CAD) of Alzheimer's

then assumes a key role by allowing to automatize its diagnosis and prevent it from being affected by inter- and intra-rater reliability from the clinicians, thus allowing for a higher accuracy to be attained both for research purposes and desirably for eventual introduction in the clinical setting.

Nonetheless, computer aided diagnosis of AD is not a straightforward procedure, as several processing steps are required for enabling the acquired brain images to be used this way. One component that is usually indispensable for this purpose consists of image registration to a standard spatial coordinate system, so that each voxel (the 3D equivalent of an image pixel) corresponds to the exact same anatomical structure across all subjects and all imaging modalities (for a multi-modal approach) [10]. However, this process, and particularly deformable image registration, can be troublesome, facing various challenges such as inter-subject anatomical variability, intensity inhomogeneity, background noise, low contrast, protocol differences, pathology-induced missing correspondences, amongst others [10]. On one hand, this is a time consuming step that makes computer aided diagnosis unpractical, having indeed led some research efforts to be placed in turning it more efficient [11]. On the other hand, many image registration algorithms are available, adopting different deformation models, similarity and regularization strategies, naturally resulting in very distinct outcomes and values of registration accuracy, each outperforming others when facing some specific registration challenge, largely depending on the application [10].

By applying the image registration step prior to image classification for AD diagnosis one can thus indirectly introduce misclassification errors, namely when there is prior knowledge of which are the regions of interest (ROI) for the diagnosis, as is the case in AD [12], and this information is affected by a poor coordinate transformation to the standard space. Considering the aforementioned disadvantages and the global burden of Alzheimer's disease, the possibility of using a method that is robust to registration errors or that can be applied independently of this step was, thus, the motivation for the current thesis.

1.2 Objectives

The goal of this thesis is to develop a supervised machine learning tool for computer aided diagnosis of Alzheimer's disease that doesn't require image registration. Ideally, this method should allow for an accurate classification of the neuroimaging data completely independently of this step. However, the idea of applying it only for training the classifier, but not on the test set, is also of interest and thus considered within the framework of this thesis.

In the attempt to develop such a method that is less time consuming and that presents an accuracy comparable to that of state of the art tools that use image registration, different approaches are explored and implemented. By applying the proposed methods to both registered and non-registered brain images, conclusions regarding efficiency, accuracy and robustness of these tools can then be drawn, achieving the objectives of this work.

1.3 Thesis Outline

This thesis is divided into six main chapters. In the second chapter, clinical background on Alzheimer's disease is presented. This includes its symptoms, etiopathogenesis, diagnosis, considerations regarding the early stages of the disease, its treatment and management. Some focus is given to the role of positron emission tomography (PET) in diagnosis, since this is the imaging modality used in the experimental part of this thesis.

In the third chapter, a brief presentation of the state of the art techniques on the subject of computer aided diagnosis of Alzheimer's disease is given. Being a hot research topic, an exhaustive description of the methods proposed in the literature for this computer aided diagnosis becomes impossible. The referred chapter will, thus, aim at briefly explaining which neuroimaging modalities have attracted more interest, whether the image registration step has been disregarded in any recent studies, which features have been used for data classification and which sort of classifiers have proven to be more accurate. Although the topic of deep learning goes beyond the explicit range of this thesis, since it has proven to perform extremely well in a broad variety of areas including, precisely, computer aided diagnosis of AD [13], being even applied in this thesis and compared against the remaining proposed supervised machine learning tools that constitute its major focus, a brief description of its recently used architectures for this purpose and of the results that followed is also included.

In the fourth chapter, inspired by the referred state of the art techniques presented in the literature, the methods implemented in this study to perform the classification task while attempting to avoid image registration are fully explained. The texton-based approach [12] used in the development of this thesis is thus introduced and its underlying mathematical principles comprehensively described. Subsequently, the analyzed datasets (including synthetically generated images from affine transforms on a registered data collection) are characterized, and the building blocks used for code implementation clarified. The developed computational framework is, at that point, fully explained. In line with what was mentioned above, the mathematical framework and code implementation behind the deep learning strategy that was applied in this thesis, a stacked sparse autoencoder [14], is also presented.

In the fifth chapter, the results from applying the different methods explored are presented and discussed. By testing them across images with varying degrees of alignment, conclusions are drawn regarding accuracy and robustness. These results refer to two binary classification tasks, one between cognitively normal and Alzheimer's disease bearing subjects, and the remaining, more difficult but of even superior clinical relevance, relative to the diagnosis of an early prodromal stage of the disease, particularly between healthy (control) subjects and with mild cognitive impairment.

In the sixth chapter, a few final remarks concerning this work are made, and indications regarding possible future developments are given.

Chapter 2

Clinical Background

2.1 Introduction

First described in 1906, Alzheimer's is a chronic progressive neurodegenerative disease, ultimately fatal, and the most common cause of dementia, accounting for around 60-80% of all its diagnosed cases [6],[15],[16],[17]. As a form of dementia, its characteristic symptoms, resulting from damage or destruction to neurons in brain regions involved in cognitive function, are difficulties with memory, language, problem-solving and other cognitive skills that affect the ability to perform everyday activities [1],[2]. Nevertheless, symptoms vary among people with Alzheimer's dementia, and the differences between its early signs and typical age-related cognitive changes can be subtle [2].

The most common early symptom in AD is in fact difficulty in remembering recently learned information, and as the disease advances, from mild to moderate to severe at a pace that varies from person to person, its symptoms start to include disorientation, deepening confusion about events, loss of motivation, failure to perform self-care, and finally the loss of bodily functions leading to death [1],[16].

In the mild stage, most people can function independently in many areas, being able to work and participate in their usual activities, but are likely to require assistance to maximize independence and remain safe. Entering the moderate stage (in some cases the longest), they may start having difficulty performing routine tasks, begin wandering, and start having personality and behavioral changes, like suspiciousness, confusion and agitation. In the severe stage, the effects of the disease become especially apparent, as patients start requiring help with basic activities and exhibiting limited ability to verbally communicate. At this point, entering the final stage of the disease, the damage to areas of the brain involved in movement and swallowing can cause individuals to become bed-bound and require continuous care, becoming vulnerable to blood clots, skin infections, sepsis (which can result in organ failure) and infections from particle deposition in the trachea, yet another contributing cause of death among many individuals with Alzheimer's [1].

Having been recognized as a common cause of dementia and a major cause of death, AD has become a well-established focus of scientific research. However, a lot is yet to be clarified, as there is still no cure for the disease and many details regarding its etiopathogenesis are yet to be fully understood.

2.2 Etiopathogenesis

2.2.1 Etiology and risk factors

AD is a multifactorial, genetically complex, and heterogeneous disease with two distinct categories, namely early onset familial Alzheimer's disease (FAD), with well-defined genetic causes, and late onset sporadic Alzheimer's disease (SAD), accounting for over 95% of diagnosed cases and where age is the dominant risk factor. Dementia onset due to the accumulating AD lesions is at around 40 years for patients with Down's Syndrome (DS), 40–60 years for FAD and over 65 years for SAD, with lower onset ages being associated with shorter life spans [17]. Not only identical AD brain lesions are present in SAD and FAD, but can also be found in DS and frequently in the elderly non-demented [17]. This suggests that it may be a result of aging, which in turn can lead to symptom manifestation depending on whether clinically significant levels have been attained [2].

Regarding FAD, three genes have been identified as potential causal factors, with most cases being associated with mutations in the one for protein presenilin 1 (PS1), some for presenilin 2 (PS2) and a few for the amyloid precursor protein (APP), respectively in chromosomes 14, 1 and 21 [2],[17]. While individuals inheriting the former and latter are guaranteed to develop the disease, those inheriting the mutation in the gene for PS2 have a 95% chance [1].

In SAD, on the other hand, such causal factors haven't been identified. Instead, it appears to arise from a combination of genetic and environmental risk factors and aging, the major risk factor genes being the e4 allele of apolipoprotein E (*APOE* e4), the best-known one, sortilin-related receptor 1 (*SOLR1*), among others [7]. Regarding *APOE*, it's located on chromosome 19 and while three gene forms exist (associated to e2, e3 or e4), inheritance of the e4 allele is the one associated with a higher risk for AD, as, compared to the possession of two copies of the e3 form (for which the risk is still higher than for e2), having one copy of the e4 allele leads to a threefold risk, and the homozygous condition brings an eight-to-twelve fold risk of developing AD [1],[2]. Still, different studies have found that only about 65% of patients diagnosed with Alzheimer's in the USA presented at least one copy of the e4 allele [1], such that its inheritance, among other risk genes, doesn't guarantee that an individual will develop the disease. The remaining greatest risk factors consist of age and family history, as individuals with one or more first-degree relatives with AD are more likely to develop it, as well as others including cerebrovascular disease, hypertension, diabetes, dyslipidemia, traumatic brain injury, depression, cancer, as well as low physical and cognitive activity [1]. Since a vast portion of these are modifiable, they can be taken into account in preventing Alzheimer's; moreover, this corroborates the importance of achieving an accurate diagnosis at an early stage, allowing for a proper management so as to correct the behaviors associated with those risk factors and prevent more severe stages of the dementia to be entered.

As will be explained in 2.2.2, a major risk of developing AD is also present in individuals with Down's Syndrome, who possess an additional (third) copy of chromosome 21. Besides, as with all adults, advancing age increases the likelihood for Alzheimer's, such that about 30% of people with Down syndrome who are in their 50s have Alzheimer's dementia and 50% or more will develop it in their lifetimes [1].

2.2.2 Pathogenesis

To the present date, two hallmarks of Alzheimer's disease regarding major structural brain changes have been established, particularly plaques and neurofibrilliary tangles, usually co-localized with neuronal and synaptic loss [6],[7],[17],[18]. With respect to the former, it occurs due to the accumulation of beta-amyloid ($A\beta$) peptides outside neurons, and it's believed to contribute to cell death by interfering with neuron-to-neuron communication at synapses [1]. Regarding the latter, it consists mainly of deposits of an abnormal (hyperphosphorylated) form of protein tau inside neurons, that causes the transport of nutrients and other essential molecules there to be blocked [1]. As mentioned, the two are a normal part of aging, presumably since it's accompanied by a progressive increase in oxidative stress.

Indeed, concerning neurofibrilliary tangles, it's believed that the increase in oxidative stress causes an elevation of calcium concentration in the intracellular compartments, which in turn activates calcium-dependent catabolic enzymes, as kinases, that progressively phosphorylate tau and decrease its binding strength with microtubules. Phosphorylated tau can then self-assemble and form paired helical filaments and, in turn, neurofibrillary tangles, while the destabilized microtubules break down [17]. Neurofibrillary tangles form in some cell bodies whose axons terminate in regions containing abundant amyloid-bearing neuritic plaques, suggesting they may be related to amyloid plaque formation in AD brain; however, they also occur in numerous etiologically distinct neurological disorders that show no amyloid deposition, such that these can be viewed as a somewhat nonspecific response by certain central nervous system neurons to a variety of insults [18].

Amyloid (neuritic) plaques, on the other hand, consist of a central deposit of extracellular amyloid fibrils, surrounded by dystrophic neurites (both dendrites and axonal terminals), activated microglia, and fibrillary astrocytes, and although plaques with these characteristics can occasionally be observed in other age-related degenerative brain diseases, they occur abundantly in three conditions, namely, AD, Down's Syndrome, and, to a lesser extent, normal brain aging [18].

It's believed that the excessive deposition of $A\beta$ leading to plaque formation in AD, and in particular of the disease-associated $A\beta42$ isoform, derives from proteolytical processing of the amyloid precursor protein following the amyloidogenic pathway [16]. Amongst other mechanisms yet to be clarified, this results from an up-regulation of *APP* transcription, causing an increased biosynthesis of one or more isoforms and an excess number of precursor molecules for proteolytic processing [16]. Some of these will then undergo alternative cleavage at or near the N- and C-termini of the $A\beta$ fragment, instead of the usual constitutive one within that region, which would preclude the formation of $A\beta$ fragments [16]. This alternative proteolytic pathway also occurs in healthy subjects, as corroborated by the development of some $A\beta$ deposits during aging, but in a lower frequency compared to the cases of FAD and Down's syndrome, which ultimately exhibit similar histological phenotypes [18]. Indeed, since chromosome 21 includes the gene for APP, the additional copy present in DS makes it a high-risk factor for AD, as an increasing number of fragments (or more aggregation-prone) can be synthesized and accumulated, leading to a much earlier development of the AD-type pathology [1].

Moreover, it has been shown that amorphous, largely nonfilamentous deposits of $A\beta$ that have little

to no surrounding dystrophic neurites or glia ("diffuse" or "preamyloid" plaques) are even more common in AD than neuritic plaques, even in brain regions that appear to be largely clinically unaffected [18]. While the proximity between amyloid cores and degenerating neurites (including axonal terminals) and activated glial cells could suggest that the extracellular amyloid derived from its accumulation in altered neurons, astrocytes, or microglia, studies in DS and AD indicate that degenerating neurites and glia are not a prerequisite for $A\beta$ deposition [16]. The formation of the more compact neuritic plaques then results from the conversion of the fibril form of the peptides present in these plaques with advancing age, again owing in part to increasing oxidative stress [17]. Indeed, once this exceeds a given threshold in regions with high metabolic rates, microglial cells are activated and able to catalyze the oxidative conversion of preamyloid to amyloid, as well as initiate a chronic inflammation and multicellular degenerative response, with neuronal and profound synaptic loss in selected brain regions, resulting in brain atrophy [1],[17]. This chronic inflammation can in fact induce a mixture of trophic and toxic effects on surrounding neurites, leading to cytoskeleton alterations that can account for how neurites of varying transmitter specificities can be found in a single plaque, as well as, in some cases, for the development of tau-containing paired helical filaments in the cell bodies of some neurons whose axons project to amyloid-bearing plaques [16]. Regarding the oxidative conversion from preamyloid to amyloid, it can be further catalyzed by the reaction of ApoE4 with O2, constituting a risk factor for AD, as stated [17].

Moreover, despite the hypothesis that secreted $A\beta$ gradually increases in the extracellular space until it starts to form aggregates of insoluble β-pleated amyloid fibrils and exert its toxic effects there, the intermediate products of $A\beta$ fibril formation (namely lower molecular weight oligomers and protofibrils) have recently been suggested to be more toxic and disruptive to synaptic plasticity than fibrils, while soluble forms of $A\beta$ should also correlate better with cognitive function in AD than insoluble ones [16]. Thus, although this hypothesis may not yet have as wide support as the established extracellular one, there is evidence that intraneuronal accumulation of $A\beta$ peptides plays an important role in the development of AD, such that both may be significant in its pathogenesis [16]. Indeed, it was previously reported that extracellular $A\beta 42$ led to some internalization by neurons, and this can accelerate the aggregation of intracellular $A\beta$ after neuritic degeneration; at the same time, degenerating neurites and synapses may release newly generated intracellular $A\beta$ into the extracellular spaces, which may then be propagated to the adjacent neuronal processes, and so the two hypotheses can be interconnected [16].

Regarding neurotransmitter alterations, deficits in hippocampal and cortical choline acetyltransferase (the enzyme responsible for acetylcholine syntesis) and shrinkage and loss of cholinergic projection neurons in the basal forebrain have been documented, so that, as will be presented in section 2.5, most currently approved pharmacological treatments are acetylcholinesterase inhibitors, which, preventing acetylcholinesterase (AChE) from breaking down acetylcholine, increase its concentrations at cholinergic synapses [1],[2],[18],[19]. Although the reported deficit must contribute to the progressive memory impairment associated with the disease, other neurotransmitters (namely norepinephrine, serotonin and dopamine) and neuromodulators are also affected in AD, suggesting that, unlike in other disorders such as Parkinson's (regarding dopamine), the clinical symptoms in AD don't reflect a deficit in a single neurotransmitter system, leading most research efforts to be placed in the two aforementioned lesions [18].

2.3 Diagnosis

2.3.1 Definitive

As there is still no single and inexpensive test for Alzheimer's, a variety of tools is normally used to help make a clinical diagnosis and exclude other potential causes of dementia, including brain tumors, Parkinson's disease, multi-infarct dementia, Huntington's disease, Progressive Supranuclear Palsy, normal pressure hydrocephalus, subdural hematoma, multiple sclerosis, history of significant head trauma followed by persistent neurological deficits and known structural brain abnormalities, low thyroid function, vitamin B12 deficiency, infections, addiction, cancer, depression, among others [1]. To establish this differential diagnosis, medical and family history, cognitive tests and physical and neurologic examinations, blood tests and different neuroimaging modalities can be applied [1]. However, this diagnosis can be unreliable, with an overall accuracy as judged by autopsy confirmation close to 78% and even lower (around 50% to 60%, according to different studies) for early AD, since several of its symptoms are shared with other neurological disorders [7].

While the increasing disease severity can support a clinical diagnosis, a definitive one is only possible at autopsy, in the presence of characteristic AD pathologic brain lesions (amyloid plaques and neurofibrillary tangles). Then, to assess neurofibrillary changes, the most commonly used measure is the Braak score, rating its distribution on silver-stained sections from various neocortical regions, the entorhinal cortex, and hippocampus, and reported to correlate with dementia severity (and to some degree with cerebral atrophy), recognizing six different stages of AD [7]. Another criterion, by the Consortium to Establish a Registry for Alzheimer's Disease (CERAD), is based on the semiquantitative assessment of the density of neuritic plaques on silver-stained sections from various neocortical regions, also adjusting the resulting score for age [7]. Most structured guidelines for clinical–pathological correlation of AD consider, thus, both the CERAD neuritic plaque score and the neurofibrillary Braak score, to express the likelihood that the lesions account for the dementia symptoms of the patient, while neuropathological changes based on $A\beta$ plaque score can also be also considered [7].

2.3.2 Neurological assessment

As previously mentioned, given the cognitive and functional impairments that Alzheimer's disease patients present, its diagnosis comprises performing neuropsychological assessment tests, most of which are based on attention, calculation, comprehension, construction, naming, orientation, recall, registration, repetition, spelling, and writing. Among those, the most commonly performed in the clinical setting include the Mini-Mental State Examination (MMSE, the most widely used and well validated one), the Clock Drawing Test (CDT), Alzheimer's Disease Assessment Scale (ADAS), the Clinical Dementia Rating (CDR), the Global Dementia Scale (GDS) and the Neuropsychological Test Battery (NTB) [7], from which MMSE and CDR will be explained in more detail, as they are the ones considered in the experimental part of this thesis. While the available cognitive scales provide different values of sensitivity and specificity in distinguishing between types of dementia and different stages of Alzheimer's disease,

some enabling to discriminate between cognitively normal and impaired subjects, or between moderate to severe cases of AD, and even though combining two or more of these tests can improve detection of early stages of the disease, namely of MCI (as is being evaluated for the combination of CDT and MMSE), none can be sufficiently accurate and reliable for early stage AD detection, which would be the desired stage for therapeutic intervention [7].

While in MMSE a lower score is observed in more severe stages, in CDR these are associated with higher scores. The former examines seven different modalities of cognitive ability (namely orientation in place, orientation in time, short-term verbal recall, immediate recall, language ability, numerical calculation ability and figure construction ability) through a 30 items questionnaire, resulting in a 0 to 30 scale, with a suggested cut-off score for dementia of 26 [15]. Besides requiring reading and writing capabilities, it also requires hearing, which constitutes a disadvantage of the test; furthermore, while it can be effective in differentiating cognitively normal and impaired cases, providing a sensitivity and specificity of about 79% and 95%, respectively, it's limited in differentiating between types of dementia [7].

Regarding CDR, it examines six different cognitive and behavioral domains through a semi-structured interview, comprising a set of questions for the subject and another for the caregiver. The former examines memory, orientation, judgement and problem-solving ability, while the latter, directed at the informant, also refers to the subject's personal life and hobbies. The test is based on a 0 to 3 scale, where 0 corresponds to no dementia, 0.5 to questionable or very mild (as later revised [20]), 1 to mild, 2 to moderate, and 3 to severe dementia [21], with the designation of MCI being often supported by a global rating of 0.5 on this scale [19], although some studies suggest that this can be indicative of a "pre-MCI" stage [20]. Although CDR can result in a moderate to high inter-rater reliability, early dementia detection still presents limitations; for this purpose, a modified CDR (mCDR) has been introduced to diagnose and stage MCI and was found to reliably distinguish between MCI and subjects with normal cognition, with the disadvantage of not including objective cognitive testing, thus not being able to distinguish between different forms of dementia [22].

2.3.3 Neuroimaging tools and biomarkers

The inherent uncertainty in a clinical diagnosis of AD, as mentioned in the previous sections, has driven a search for biomarkers, defined as "an objectively measurable substance, characteristic, or other parameter of a biological process that enables assessment of disease risk or prognosis and provides guidance for diagnosis or monitoring of treatment" [6]. In this context, brain imaging assumes a key role and one modality that has been recognized over the last decades for this purpose is FDG-PET [9].

Indeed, as the brain relies almost exclusively on glucose as its source of energy, the glucose analog FDG is a suitable indicator of its metabolism and, when labeled with the radioactive isotope Fluorine-18, the resulting 2-[^{18}F]-fluoro-2-deoxy-D-glucose is conveniently detected with PET [9],[23]. Although this modality is relatively expensive, requires intravenous access and involves exposure to radioactivity (even if at levels well below significant known risk), its application for the measurement of regional cerebral glucose metabolism (rCMRglc), a marker of synaptic activity, has become a standard technique in both oncology and dementia research, being technically mature and well tolerated by patients [9],[24],[25].

With FDG-PET, the metabolic abnormalities that occur in AD can thus be properly identified, which are assumed to result from a combination of processes putatively involved in its pathogenesis including specific gene expression, mitochondrial dysfunction, oxidative stress, deranged plasticity, excitotoxicity, glial activation and inflammation, synapse loss and cell death [9].

On one hand, FDG-PET can be used to exclude potentially surgically treatable causes of cognitive decline, or identify non-AD (or even mixed) dementias, namely Dementia with Lewy bodies (DLB) or frontotemporal dementia (FTD) (where hypometabolism in the occipital region along with the temporoparietal would support DLB, and the prominent presence of frontotemporal rather than temporoparietal hypometabolism indicate FTD [9],[26]). On the other hand, it can corroborate a clinical diagnosis of AD in symptomatic individuals and even constitute a robust biomarker of neurodegeneration, where hypometabolism can be observed to precede the appearance of cognitive symptoms, and predict the rate of cognitive decline in individuals who progress to AD with high sensitivity (namely in patients carrying risk factors such as the *APOE* e4 allele or others diagnosed with MCI) [9].

So, despite providing a nonspecific indicator of metabolism that can be altered for a variety of reasons (namely ischemia or inflammation), possibly irrelevant or only indirectly related to AD, a characteristic ensemble of limbic and association regions that are typically hypometabolic in clinically established AD patients has been identified [9]. Such reductions in rCMRglc are found in neocortical association areas including the posterior cingulate, precuneus, temporoparietal and frontal multimodal association regions, as well as the hippocampus and medial temporal cortices, as illustrated in Figure 2.1 (retrieved from [9]), while, and in contrast to other dementia types, the primary visual cortex, sensorimotor cortex, basal ganglia and cerebellum are relatively unaffected [24],[26]. These observations generally reflect AD clinical symptoms, with impairment of memory and associative thinking (including higher-order sensory processing and planning of action) but relative preservation of primary motor and sensory function [24]. The anatomical distribution of functional changes is indeed closely related to impairment of some localized specific cognitive functions and their severity increases with progression of the disease, while there may also be a distinct hemispheric asymmetry (more common in early stages) corresponding to the predominant cognitive deficits (language impairment in the dominant and visuospatial disorientation in the subdominant hemisphere) [24]. Comparing to other neuroimaging modalities, the abnormalities found in AD with FDG-PET mirror those found with SPECT and MRI, also being reported to be more reliable for diagnostic purposes than these [24].

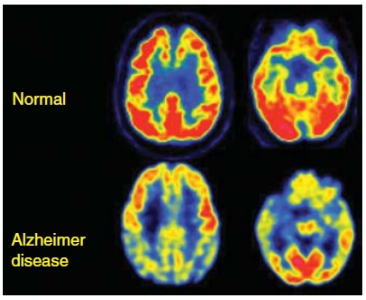


Figure 2.1: Example of transaxial FDG-PET images of cognitively normal vs. mild AD subjects.

Less severe or consistent hypometabolism has been identified in MCI patients, some of whom later convert to AD, such that it parallels cognitive function along the trajectory of normal, preclinical, prodromal, and established AD [9]. Also, even though vascular disorders (namely ischemia, amyloid angiopathy and micro-hemorrhage) potentially confound the relation between the FDG pattern and the clinical phenotype, it remains well correlated with the histopathologic diagnosis of AD at autopsy [9].

Since at present brain PET scans are often interpreted in a qualitative manner only by visual reading, which heavily depends on observer experience and training and lacks a clearly defined cutoff to distinguish between normal and pathological findings, in order to be well suited as a biological marker for AD, FDG-PET should be complemented by an objective image analysis procedure that can be widely and easily applied in different PET centers – section 3 will go into more detail in this aspect, regarding computer aided diagnosis of Alzheimer's disease. Apart from the glucose metabolism as shown by FDG-PET, constituting a biomarker for neurodegeneration, other factors are also being studied as possible biomarkers for AD. For instance, beta-amyloid PET imaging can accurately reflect the presence of neuritic plaques in the brain, such that elevated levels of its measured $A\beta$ retention give clinicians reason to conduct additional Alzheimer's testing (and in the case of MCI also greatly increase the likelihood of it being due to AD), while non-elevated levels indicate a reduced likelihood that cognitive impairment is due to Alzheimer's [9],[27]. Among the radiotracers used for this purpose are ^{11}C-labelled thioflavin analogue, named for convenience Pittsburgh compound B (^{11}C-PIB) and fluorinated tracers such as florbetapir, flutemetamol and florbetaben [1],[23]. In comparison with FDG-PET, longitudinal data has shown that, once the stage of established AD is reached, amyloid deposition in most regions has plateaued, while FDG hypometabolism continues to accentuate along with the decline in cognitive function and is accompanied by decreases in neurotransmitter levels and brain volume, and so while PET amyloid imaging could be a sensitive tool for the early detection of prodromal AD, measurement of cerebral glucose metabolism by ^{18}F-FDG PET will provide more information about disease progression [23]. Moreover, several groups have observed high amyloid deposition in parietal regions in association with co-localized FDG hypometabolism, possibly indicating a local toxicity; nonetheless amyloid deposition is also commonly reported in the frontal cortex in Alzheimer's disease, suggesting that these are separable phenomena [9].

Structural MRI can be used to detect cerebral atrophy, which is also being explored as a possible biomarker for neurodegeneration and neuronal injury [1]. As FDG hypometabolism, brain volume loss has also been reported in cognitively normal individuals who go on to develop AD, and, although it is also observed in AD hypometabolic areas, studies have suggested that, once again, this and function loss are detachable phenomena in AD [9].

Regarding neurofibrillary tangles, elevated cortical tau via PET imaging also constitutes a possible biomarker, for which the tracers should have high affinity to hyperphosphorylated tau paired helical filaments, and high selectivity for it over $A\beta$ and other protein deposits [28]. Several tau PET tracers including T807, THK-5117, and PBB3 have in fact been developed and succeeded in imaging neurofibrillary pathology in vivo, although further studies are required to evaluate their reliability and quantitative performance, and to validate their in vivo binding selectivity to tau pathology [28].

Additional biomarkers currently being studied for Alzheimer's disease are found in CSF and blood, which, despite reflecting the rates of both protein production and clearance at one point in time rather than the cumulative damage assessed by neuroimaging biomarkers, may provide insight into the pathological changes of Alzheimer's [1]. These include CSF $A\beta 42$ (for which lower levels are indicative of beta-amyloid deposition in the brain) and CSF phosphorylated tau and total tau (for which higher levels are biomarkers of neurofibrillary tangles and neurodegeneration) [1].

Advanced MRI techniques such as diffusion tensor imaging (DTI) and associated tractography technologies, arterial spin labeling measures of cerebral blood flow and PET tracers targeted at the cholinergic system, microglial activation and other tracers in development, are also contributing to a better understanding of AD [9]. Although these biomarkers of functional impairment, neuronal loss, and protein deposition that can be assessed by neuroimaging (structural and functional MRI and FDG, tau and amyloid PET) or CSF analysis are increasingly being used to diagnose Alzheimer's disease in research studies and specialist clinical settings, its clinical usefulness still requires extensive validation, as stated [1],[6],[9]. Introducing these in the diagnosis routine would not only largely improve its accuracy, but also allow for detecting preclinical stages of AD, in which the following section will go into more detail.

2.4 Early Stages

2.4.1 Mild cognitive impairment

According to the 2011 revised guidelines for clinical diagnosis of AD, it's considered that the disease begins when the associated brain changes start to occur, possibly 20 or more years prior to the manifestation of symptoms of dementia [1],[7]. Since at that point they can still be compensated for, individuals continue to function normally; however, with the advance of neuronal damage, cognitive decline begins to show as the mild to moderate to severe stages of the disease are entered [1]. Two stages of early AD prior to symptoms manifestation are thus recognized, namely preclinical Alzheimer's disease and MCI, each accompanied by the proper modifications in the disease-related biomarkers [1],[7],[15]. Figure 2.2 (retrieved from [7]) illustrates the hypothetical deviation of pathological and, later, clinical trajectories of AD from the normal aging process, as well as the ideal early diagnosis of AD which will eventually be enabled by the incorporation of well-validated biomarkers of the disease.

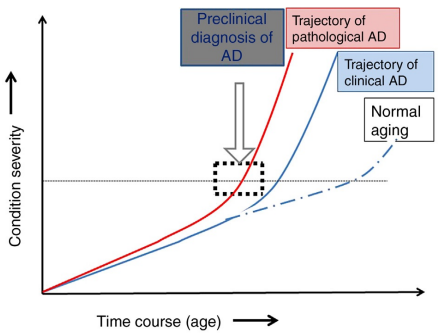


Figure 2.2: Depiction of hypothetical pathological and clinical trajectories of AD against normal aging.

Regarding MCI, where, as mentioned, despite the exhibited cognitive decline, individuals can still perform everyday activities without notable impairment, the prevalence in elderly individuals (of age 65 or older) is in the range from 15% to 20% [1],[6],[7]. This stage corresponds to the earliest clinical manifestation of AD, having heterogeneous presentations and underlying pathologies [1]. Indeed, not only it includes incipient Alzheimer's disease, but also other causes of dementia, as well as a form of cognitive impairment that doesn't progress to dementia, remaining stable, or that may even revert to normal cognition [1],[7],[15]. Since all subtypes of MCI are associated with an increased risk for AD and other dementias, the importance of older adults reporting their own experience of memory and thinking problems, without (or before) a formal examination by a doctor (named subjective cognitive decline) has been recognized, to be followed by medical help seeking for diagnosis and possible treatment [1]. Indeed, as evidence suggests that up to 33% of Alzheimer's disease cases are attributable to modifiable risk factors, early detection of MCI is important, even if it outpaces the therapeutic options, as it can still have several clinically significant implications, like slowing the rate of cognitive decline, reducing morbidity-affected life years, and improving quality of life [20].

MCI can be divided into two categories: amnestic MCI (aMCI) and multimodal MCI (mMCI) or non-amnestic MCI [7],[8],[15]. MCI with primarily memory deficits is called amnestic MCI, where the memory complaint should be corroborated by an objective examination [7],[8],[15]. Multimodal MCI, on the other hand, is identified as multiple or isolated extra-memory cognitive impairments, namely regarding thinking skills, inability to make sound decisions and judgments, and inability to take the sequential steps when performing complex tasks [7],[8],[15]. In general, individuals with aMCI eventually develop AD and those with mMCI develop non-AD dementias [7].

In fact, regarding aMCI, studies suggest that up to 67% of patients have underlying Alzheimer's pathology, 15% to 25% have neurodegenerative diseases other than Alzheimer's disease (such as frontotemporal degeneration or Lewy body disease, among others), and the remainder have normal age-related changes [6]. Regarding MCI to AD conversion, studies have found that the annual rate ranges from around 10% to 25%, and that within 5 years the conversion is expected to conclude [7]. Other studies have found that this conversion at 5 years follow-up took place in around 32% of patients diagnosed with MCI, with that for dementia rounding 38% [1]. Patients with MCI who progress to AD have converter MCI (c-MCI), and those who don't have stable MCI (s-MCI) [7]. According to the previous section, MCI symptoms accompanied by elevated levels of β-amyloid are indicative of an early stage of Alzheimer's (called MCI due to Alzheimer's disease), and similar conclusions can be taken for FDG hypometabolism [1],[8],[23]. Thus, a combination of neuropsychological testing (namely CDR, frequently used to assist in the evaluation of MCI) and neuroimaging improves the diagnostic accuracy of predicting cognitive decline in people in this phase compared with that achieved with either modality alone [8].

2.4.2 Preclinical Alzheimer's disease

Regarding a preclinical state of AD, preceding MCI, its existence is still speculative and has yet to be proven, although its evidence includes the presence of genetic risk factors, AD-like brain images, and abnormal CSF biomarkers in cognitively normal individuals, and AD-like neuropathology seen in the

normal aged brain [1]. On the basis of the extent of amyloidosis and neurodegeneration, recent guidelines suggest that preclinical AD can be categorized into different stages, the last of which including subtle cognitive decline that has not reached MCI levels [7]. Those cases with no evidence of amyloidosis (amyloid negative) but signs of neurodegeneration (neurodegeneration positive) are categorized as suspected non-Alzheimer's pathophysiology, while individuals greater than 65 years of age with intact cognitive ability before death and substantial AD lesions at autopsy are classified as having asymptomatic or preclinical AD, the phase during which the diagnosis would indeed provide a critical window of opportunity for therapeutic intervention [7].

2.5 Treatment and Management

A cure for Alzheimer's disease is yet to be discovered. Indeed, even though there are medications available today for this disorder, none of those pharmacologic treatments can slow or stop the neuronal damage and loss behind its symptoms, which make the disease fatal [1],[27].

There are currently six FDA-approved drugs for its treatment, namely rivastigmine, galantamine, donepezil (the most commonly prescribed), memantine, memantine combined with donepezil, and tacrine [1]. Apart from memantine (a glutamatergic partial antagonist found effective in more severe dementia), these are cholinesterase inhibitors, which, as mentioned, putatively inhibit AChE causing a transient increase in acetylcholine levels at the synapse [2],[15],[23],[27]. Several reviews show that this class of medications has a moderate but worthwhile symptom modifying effect in Alzheimer's disease so that, despite not being curative, it can indeed temporarily (during 6 to 18 months, varying interpersonally) improve cognition and/or function, after which the symptoms will re-start to manifest and patients will continue to decline [15],[23],[27]. These are also generally well tolerated, having bradycardia as the most concerning side effect, and the most usual ones being gastrointestinal upset (which can be precluded by opting for a rivastigmine transdermal patch) and sleep disturbances [15],[27].

Non-pharmacologic therapies are also often used with the goal of maintaining or improving cognitive function, the ability to perform activities of daily living or overall quality of life, thus causing a significant positive impact in patient's lives, particularly exercise and cognitive stimulation [1],[2],[15]. Moreover, computerized memory training, listening to favorite music to stir recall, and incorporating special lighting to lessen sleep disorders, may be used to reduce behavioral symptoms like depression, apathy, wandering, agitation and aggression [1]. Nonetheless, as with current pharmacological therapies, these do not slow or stop the progression of AD, having benefits to cognitive function last only up to 3 months [1],[2].

The fact that neither of these therapies can serve as a cure for Alzheimer's thus corroborates the importance of achieving an accurate clinical diagnosis still in its early stages (allowing for a proper management of the disease), where computer aided diagnosis can play a key role.

Chapter 3

State of the Art

3.1 Neuroimaging Modalities

As discussed in the previous section, neuroimaging data provides valuable insight for the diagnosis of AD and related disorders (such as mild cognitive impairment) and particularly for the computer aided diagnosis system. Although the current thesis focuses solely on FDG-PET (also widely used in the literature, particularly in [12],[29],[30]), other neuroimaging modalities, both functional and structural, have been explored for the purpose of CAD of Alzheimer's disease.

One of these techniques consists of structural MRI (sMRI) which, as mentioned in 2.3.3, provides visual information regarding the macroscopic tissue atrophy in the form of a structural image of the brain, having been applied in several studies such as [13],[31],[32],[33],[34],[35],[36],[37],[38],[39],[40],[41], [42],[43],[44].

Neuroimaging data acquired with SPECT has also been used for this purpose, namely in [45], which, similar to FDG-PET, instead of imaging anatomical structures, provides functional information. While the former informs about cerebral glucose metabolism, by using a gamma ray emitting radionuclide and particularly using a regional cerebral flow (rCBF) agent, SPECT enables to measure cerebral blood flow as this radiotracer accumulates in regions of high rCBF.

Besides being used as the unique source of information, these neuroimaging modalities have also been combined for multi-modal approaches, both with each other and with other clinically relevant data, such as *APOE* genotype information (having the e4 allele constitute a risk factor, as mentioned in 2.2.1), or the result of neuropshychological tests. Naturally, the motivation for the combination of these approaches derives from the fact that each technique traces different biomarkers and, thus, complementary information can be retrieved by their fusion. Moreover, while it is possible to concatenate the multi-modal features for classification, some strategies have also been proposed to avoid the possible loss of information regarding data correlation and to improve the accuracy of the diagnosis in comparison to this simple concatenation method, namely using non-linear graph fusion (NGF), as performed in [46], or using deep learning strategies, as will be presented in section 3.6.

Several studies have indeed followed a multi-modal approach, either combining FDG-PET with sMRI

(in [47],[48],[49],[50]), or FDG-PET and sMRI also with either PIB-PET (in [51], which enables to measure brain $A\beta 42$ retention levels, leading up to neuritic plaques' formation) or genetic information (in [46]). Other combinations of these modalities that have been explored include the integration of sMRI with CSF features (which, as mentioned in 2.3.3, provide important clinical insight regarding $A\beta 42$ deposition in the brain, as well as of neurofibrillary tangles' formation, with the disadvantage of being an invasive procedure, requiring a lumbar punction) and neurological assessment results, namely from FAQ (which stands for Functional Activities Questionnaire) and CDT (in [52]). Moreover, sMRI has also been combined with advanced magnetic resonance imaging techniques, including DTI (in [53], which can inform about water diffusion in the brain through two DTI-derived maps, namely the fractional anisotropy (FA) and the mean diffusivity (MD), the latter of which represents its magnitude, having higher values be associated with a faster diffusion, which can accompany the neurodegenerative process as this results in a loss of barriers that would restrict the motion of water molecules in brain tissues) as well as ASL (in [54], which enables to measure brain perfusion in terms of cerebral blood flow (CBF), as mentioned).

3.2 Image Registration

As introduced in section 1.1, an interesting application in what refers to CAD of Alzheimer's disease consists of the design of a system that can perform well without image registration. Since the vast majority of studies found in the literature perform voxel-wise comparisons or investigate brain abnormalities in regions of interest (for which there is indeed some prior knowledge regarding AD), and both require non-linear warping of the original or pre-processed images to some standard space (namely following some normalization, reorientation, smoothing or linear registration procedures), image registration is typically a common practice. It is important to note, however, that even though the registration to the chosen brain atlas coordinate system usually involves an automated process to spatially transform each individual brain image into the coordinate space, this doesn't guarantee that an identical point in space corresponds to the same anatomical feature for all subjects.

Different neuroimaging analysis software programs have been used for this purpose, including SPM (which stands for statistical parametric mapping, and mostly using its versions SPM8 (in [13],[29],[30], [32],[33], [34],[36],[39],[40],[41],[45],[46],[47],[48],[49],[50], [53]) or else SPM5 (in[51])), HAMMER (which stands for hierarchical attribute matching mechanism for elastic registration, applied in [32],[33],[48], [49],[50]), FreeSurfer (in [37],[41],[42],[52]), amongst others such as Advanced Normalization Tools (ANTs) and Mindboggle (in [41]). The two most widely used spaces in the neuroscience community are, in turn, the Talairach space and the Montreal Neurological Institute (MNI) space (the latter of which being the one considered for the commonly employed SPM software) [55].

Regarding the Talairach space, it is based on a stereotaxic atlas of the human brain in which the coordinate system was constructed following the identification of given anatomical landmarks. Particularly, the Talairach coordinate space has its spatial origin defined at the anterior commissure (AC, which, similar to the posterior comissure, PC, corresponds to a white matter tract connecting the two temporal lobes of the brain hemispheres), the y-axis defined as connecting the superior aspect of the AC and

the inferior edge of the PC, the x-axis as passing through the AC and being orthogonal to the AC-PC line, and finally the z-axis as passing, in the midline plane, through the interhemispheric fissure and the AC [55]. Furthermore, to account for inter-subject differences, the authors defined a proportional grid system to align each individual brain image to the atlas. This involved segmenting the brain into 12 subvolumes, considering 2 divisions laterally (left and right, using a vertical plane parallel to the y-axis), 2 vertically (above and below the AC-PC line, using one horizontal plane passing through the x- and y-axes), and 3 in the AC-PC direction (from posterior limit to PC, from PC to AC, and from AC to anterior limit, using 2 horizontal planes, parallel to the x-axis, one passing through the z-axis and another through the PC) [55],[56]. A piecewise linear scaling method was then performed on each region, separately for each direction, providing a simple way of converting each individual brain image to the Talairach space [55]. Nonetheless, this brain atlas has faced some criticism, not only because the piecewise linear scaling method is stated to perform poorly when matching the anatomical landmark locations (especially cortical regions) to the atlas when compared to other methods that use nonlinear transformation, but mostly due to the fact that it was constructed from a single subject and thus cannot be considered as representative of the neuroanatomy of the general population [55].

To attain a better representation of the average neuroanatomy, researchers at the Montreal Neurological Institute (MNI) thus created an average brain template based on the sMRI scans from several hundred individuals, named as MNI305 [57]. To achieve this, the first step consisted of identifying a set of anatomical landmarks from 250 sMRI brain images of healthy subjects and matching them to those in the Talairach atlas, through reorientation and scaling, followed by averaging of the resulting sMRI brain images [57]. An extra 55 images were then considered, and registered to the constructed 250 atlas using an automatic linear registration method, where instead of performing this mapping according to Talairach's piecewise linear model, the whole-brain linear image similarity residual, using 9 parameters, proposed in [56], was applied. The resulting template, from averaging the registered 55 brain images with the 250 manually registered ones in the previous step, is then an approximation of the original Talairach space, although the z-coordinate is approximately +3.5 mm relative to the Talairach coordinate [57]. This resulted, thus, in the creation of the MNI305 atlas.

Another template has been proposed by the same authors that is currently the standard MNI template, being widely used for image registration, and which corresponds to the average of 152 normal sMRI scans that have been linearly matched to the MNI305 using a 9 parameter affine transform, as in the previous approach [56],[57]. This template was adopted by the International Consortium of Brain Mapping (ICBM) as an international standard, leading to its designation of ICBM152. A new version was presented later on, which corresponds to the ICBM452 template, although this is not yet widely used. Furthermore, a new, single brain, atlas has also been presented by the same research group, in this case derived from scanning the same subject for 27 times and co-registering and averaging all scans to obtain a very high detail template, which was also matched to the MNI305 space, named as Colin27 [58]. Despite the fact that, similar to the Talairach atlas, this template can't capture inter-subject anatomical variability, the fact that it presents a very high image resolution has also led it to be adopted by different groups as a stereotaxic template, namely in [38],[43],[59].

It is also important to note that the MNI templates are not perfectly matched with the Talairach standard brain due to large differences in brain shape and size between the two atlas, so that there isn't a straightforward procedure for transforming multiple subject data from one space to another, and it is even probable that the same coordinate location in the MNI space of two subjects can map to different regions in the Talairach space [55],[57]. As shown in figure 3.1 (retrieved from [57]), the MNI registered brain images are indeed slightly larger (in particular higher, deeper and longer, having lower and larger temporal lobes) than the Talairach brain, and these differences accentuate further in the lateral regions in contrast to the medial ones [55],[57].

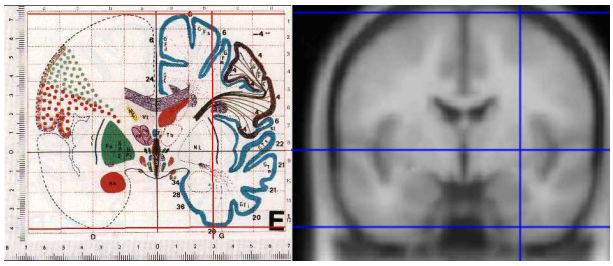


Figure 3.1: Comparison between equivalent brain slices on the Talairach and MNI space. The Talairach brain atlas is shown on the left, while the right image corresponds to the ICBM152 template.

Despite the advantage provided by the MNI coordinate space in terms of no longer being a single brain atlas, the Talairach coordinate system has, however, become the standard reference for reporting the brain locations due to its construction from a set of anatomical landmarks, so that it is still also widely used in the literature. Amongst the studies in CAD of Alzheimer's Disease that have performed image registration through non-linear warping to the Talairach brain atlas are [12],[31],[35],[42],[52], while the stereotaxic template provided by the MNI coordinate space (and particularly, with regards to the ICBM152 template) has also been used in multiple studies, including [13],[29],[30],[32],[33],[34],[36], [39],[40],[41],[45],[46],[47],[48],[49],[50],[51],[53].

Furthermore, in what concerns the state of the art as presented in this chapter (bearing in mind, however, that the studies presented here are meant to illustrate the typical procedures used, but naturally cannot cover all the work that has been performed in recent years in this subject) it can be noted that from a set of 29 research papers, summarized in 3.7 and further discussed throughout this chapter, only 4, performed in recent years, did not apply the image registration step ([44]), or only did so for training the classifier for the CAD system ([38],[43],[59]). In order to avoid this processing step, in [38], the authors opted for a landmark-based feature extraction method for fast AD diagnosis, both without nonlinear registration (in the testing stage) and tissue segmentation. For that purpose, in the training stage, a number of landmark points were detected on the template image (following non-linear registration) and a number of other (active) landmark points was added based on the similarity and inconsistency maps of the brain, and, using their respective deformation fields, these landmarks were then directly projected to the linearly-aligned training images. Its morphological features were extracted to train a linear SVM classifier for AD diagnosis [38], and a shape-constrained random forest was used to learn a non-linear mapping between the area surrounding the different voxels (patches) and its 3D displacement to the detected target landmarks. Having learned this random forest model, the testing stage could then be

executed without non-linear image registration, by using the model to estimate a 3D displacement from every voxel (based on its local morphological features) to the potential landmark position, and further aggregation of all votes to obtain a voting map from which the landmark position could be easily identified. This landmark-based method was also further extended for analysis of longitudinal MR images in [59], following a similar approach. In [43], a similar method regarding landmark detection was also applied, but, contrary to the previous studies, the classification stage was performed using a bag of patches approach, where the feature representation was further obtained from a multi-instance convolutional neural network from the input patches, each centered at each identified landmark.

In [44], on the other hand, the non-linear image registration step was fully disregarded, both for the training and test stages. For this purpose, patches within regions of interest were identified (from non-registered images), and clustered using K-means. A different deep learning strategy, particularly consisting of multiple cluster DenseNets, was then used to learn the patch-level features, which were further aggregated for region-level representations. Combining the representations of different regions, the final image classification could then be attained.

In the current thesis, as mentioned in 1.2, several methods were also explored in the attempt of developing a CAD system that can also disregard the image registration step either for both training and testing, or where this processing step is only required for the training stage. Section 4.2 will go into detail regarding the implementation of the selected approaches.

3.3 Feature Extraction and Feature Selection

In order to obtain a successful computer aided diagnosis system, a crucial aspect consists of choosing the appropriate set of features to be used at its input. Considerable research efforts have indeed been placed in identifying the most discriminative features for the purpose of Alzheimer's Disease diagnosis from the neuroimaging data acquired with the aforementioned modalities.

Several feature selection techniques have been applied for this purpose, including t-tests (in [32],[33], [35],[38],[39],[43],[48]), the separation power factor (SPF, in [29], which uses information from the area under the ROC curve, also considered in [51]), the Fisher Discriminant Ratio (FDR, in [45]), genetic algortihms (in [39]), regularization techniques (including relational regularization, in [49], both between samples, features and class labels), ANOVA and incremental error methods (in [41]) and voxel-based morphometry (VBM, used to select regions of interest based on atrophy patterns captured by sMRI, in [39]). Other statistics have also been considered, such as variance and entropy (using a threshold on these, as in [38]), inter-class variance (ICV, in [35]), mutual information (MI, in [12]) and minimum redundancy maximum relevance (mRMR, in [52]). While the previous methods are mostly based on relevant statistics on the input data, other techniques have been applied. Particularly, wrapper methods have been used in [12],[52], which select the features that lead to a better performance from the learned model. Other methods include random forests (RF, in [29],[30],[38],[43],[46]), extremely randomized trees (ET, [30]) and multiple kernel learning (MKL, [30]), which are typically designated as embedded methods, as they perform feature selection and classification simultaneously, thus including an intrinsic

variable selection process for learning the model.

Dimensionality reduction has also been attained by the use of different feature extraction methods that study linear combinations of the original variables by projecting them to a lower dimensional space, as is the case with principal component analysis (PCA, in [34],[35],[42],[44]), partial least squares (PLS, in [34]) and independent component analysis on means (ICAm, in [40]), amongst other approaches such as image reduction (in [36], excluding equal intensity voxels, and by imposing a threshold on this intensity (in [34]). Despite the fact that these are able to account for combinations of the input features during the process of dimensionality reduction, which is not always the case in the previous selection algorithms, these latter techniques bring the disadvantage of leading to increased computational costs when compared to those cases.

In what concerns the type of feature, studies on this topic have opted for either voxel-based approaches or for performing feature extraction from the whole brain, regions of interest (ROI, including particular slices of the brain) or patches of the brain.

Regarding voxel-based approaches (thus, using discriminative voxels of the brain, identified following feature selection, particularly in [12],[34],[38],[39],[43],[48],[54]), while these are intuitive and can be applied without prior knowledge of which brain regions are affected by Alzheimer's Disease, a significant disadvantage comes precisely from the fact that these disregard regional information. Similarly, methods applied to the whole brain (namely in [13],[35],[36],[40],[46],[47],[50],[54]), where neither the most discriminative voxels nor regions are identified prior to the data classification, come with the same advantages, but with crucial limitations derived from the high dimensionality of the feature vectors as well as the fact that, by including non-informative nor discriminative features, the performance of the CAD system can be negatively affected.

Feature extraction from ROIs (as performed in [12],[29],[30],[31],[35],[37],[39],[41],[42],[45],[46],[49],[51],[52],[53]), on the other hand, requires the choice of regions to be studied in advance. The idea for this kind of method results from the fact that several functionally predefined regions of the brain have been identified in multiple studies as the ones that are mostly affected by AD, as presented in section 2.3.3, so that representative features can be extracted from each region. Furthermore, as with voxel-based approaches, since only fractions of the brain are used, the dimensionality of the feature vectors can be reduced. The ROIs can either be manually labeled (namely by a medical doctor), which can be a difficult, time consuming and user dependent task, or identified through the same kind of techniques used for selecting the most discriminative features in voxel-based approaches.

A recently proposed alternative consists of performing feature extraction from patches of the brain. The motivation for this method comes from the fact that the brain regions affected by Alzheimer's, among other neurodegenerative diseases, can be part of ROIs or span over multiple ROIs, so that the simpler voxel- and ROI-based approaches (as well as those using features from the whole brain) may not effectively capture the disease-related pathologies. This patch-based approach can then be seen as an intermediate option between the ROI-based approaches and the whole brain and voxel-based ones, efficiently handling both the concerns of high feature dimensionality and the sensitivity to small changes, taking into account local information and allowing to extract richer information that might help achieve a

more accurate clinical diagnosis. Moreover, this can be similar to what is performed from the perspective of a medical doctor in a clinical setting, who may search for local distinctive regions in the brain images and then combine the interpretations with neighboring ones and ultimately with the whole brain.

In line with the approach proposed in [32], different patches of the brain can be combined, namely in a hierarchical way, so that after segmenting the full image of the brain into small patches and extracting features from each individual one, these can be combined leading to the formation of mega-patches for which higher level features can be extracted and used for classification. Another alternative can be to concatenate the feature representations derived from each individual patch, as performed in [43], and form a global representation of the brain that captures the complex relationship among image patches located at multiple landmarks, in that case using a deep learning strategy as will be presented in 3.6. In [44], on the other hand, a clustering algorithm was also applied, so that the information retrieved from the patch-based feature extraction method was also combined with regional information and dimensionality reduction could also be attained.

3.4 Feature Transformation

Several features have been considered for the purpose of CAD of Alhzeimer's disease, depending on the neuroimaging modalities and feature extraction approaches considered. The most simple and direct method consists of using the voxel intensities as features for data classification. This has been the case in different studies, namely in [12],[30], where these voxel intensities referred to FDG-PET images, in [45], using SPECT, and in [34],[39], where these values were extracted from sMRI images. Furthermore, the average voxel intensity in defined regions of interest has also been considered, both from FDG-PET, PIB-PET, DTI and ASL brain scans, particularly in [46],[49],[51],[54], as well as other related features such as the entropy of the defined ROIs, as performed in [29].

Given the fact that, as mentioned, Alzheimer's Disease is characterized by significant alterations regarding the volume, shape, density and thickness of particular regions of the brain and that these are well captured by structural magnetic resonance imaging of the brain, multiple studies using these regions' highly specific characteristics as features can also be found in the literature. Since neurodegeneration in AD mainly affects the gray matter (GM), resulting in loss of this tissue, several studies have opted for using features extracted from it (although others have focused on the cerebrospinal fluid, in [53], as well as in the white matter tissue, in [54]), namely following its segmentation in ROIs or patches. These include the volume (having been used either directly, as in [41],[42],[46],[49],[52],[54]), density (in [32],[33]) and related statistics (such as its mean and standard deviation, in [33]), amongst other shape measures in defined ROIs such as cortical thickness (in [41],[42],[52]), surface area (in [41],[42]), folding index (in [42]), intrinsic curvature (in [41],[42]), travel and geodesic depth and convexity (in [41]). Since GM loss has been described particularly in the hippocampus (HC), resulting in a decrease of its volume, as well as in other structures of the mesial temporal lobe such as the entorhynal cortex (EC), shape and volume features from these, among other regions of interest, have also been considered, namely the relative volumes of the hippocampus and amygdala and the thickness of the entorhynal cortex (in [37]),

the volume and shape of the hippocampus (namely in the form of 3D Zernike descriptors, in [51]). In [31] the authors also proposed a method, entitled scoring by non-local image patch estimator (SNIPE), which could both localize the HC and EC, and also attribute a grading score and compute each region's volume, so that the two kinds of features could, in turn, be used to obtain a clinical diagnosis.

Several research groups have also studied the possibility of using an alternative approach for feature transformation, namely the use of texture descriptors of the neuroimaging data. Among these methods are, to name a few, local binary patterns (used in [60], and its extended version of local energy patterns (LEP), applied in [38]), and texton dictionary-based approaches (in [12]). Regarding the textons method, the fact that it provides a full statistical representation of the responses to a predefined set of filters turns the resulting extracted image models into powerful descriptors, having led it to raise much interest in multiple applications, including precisely for CAD of Alzheimer's Disease. Given the success of this method, the texton-based approach has also been considered for the development of the current thesis.

As mentioned in 3.1, genetic information, neuropsychiatric tests and features extracted from the CSF have also been considered for this purpose. These include the $A\beta 42$, t-tau, p-tau (in [46]) and relative ratios t-tau/$A\beta 42$ and p-tau/$A\beta 42$ (in [52]), scores from several tests such as Functional Activities Questionnaire (FAQ), Logic Memory (LM) and Auditory Verbal Learning Test (AVLT), including the information from particular trials, delayed and immediate recall (in [52]), and *APOE* genotype information (in [46]).

Other approaches have also been considered regarding feature transformation that significantly differ from the aforementioned ones. Examples of these consist of 3D displacements from different voxels towards identified landmarks of the disease (in [38]), statistics such as inter patch (in [47]) and interslice correlation and standard deviation (in [45]), the use of discrete transform techniques (in [36]) following the conversion from 3D sMRI images to 1D signals and dimensionality reduction, the eigenvalues associated to the most important eigenbrains, following PCA (in [35]), the image projections onto the representative subspaces obtained by ICAm (in [40]), as well as other bag-of-features approaches (as is the case for the textons method), namely having the visual words correspond to Gauss–Laguerre circular harmonic functions (GL-CHFs) (in [53], obtained from sMRI and DTI) among many other feature transformation methods that have been applied for the purpose of CAD of Alzheimer's Disease.

Additionally, in recent years, several deep learning strategies have emerged which provide high-level feature representations, that can be more robust in comparison to the aforementioned hand-crafted features, and thus result in improved diagnostic performances. These alternative approaches will be exemplified and given some insight in section 3.6.

3.5 Classification

Regarding the choice of the classification algorithm, as in the current work, most studies have opted for using support vector machines (SVM), which will be explained in detail in section 4.1.2.1. The choice of this algorithm has been reported to be mostly due to its robustness in high dimensional spaces, performing well under the circumstances associated to the curse of dimensionality phenomenon. Particularly, this refers to classification problems where the size of features vectors is larger than that of the

training set, which could lead to overfitting. The fact that the SVM algorithm presents a good generalization capability, by maximizing the margin between the distinct classes to be identified, hence turns it into a good choice when dealing with these problems, which are widely common in neuroimaging data classification, due to the high feature dimensionality that usually persists regardless of the adoption of feature selection and dimensionality reduction techniques. Several studies have indeed opted for using the SVM algorithm, namely [12],[29],[32],[33],[34],[35],[36],[37],[38],[39],[40],[42],[45],[48],[49],[50],[51],[52],[54].

In the same context, some studies have also opted for using multiple kernel learning (MKL) SVMs, which, despite being more computationally expensive and less intuitive in its interpretation, can perform better namely in multi-modal approaches, as it is obtained by a linear combination of kernels and it might be the case that different kernels adjust better to given sets of features but not others that would require different notions of similarity and hence different kernels. Thus, in this case, instead of building a model with a specific type of kernel separately for each modality, one can opt for the MKL approach and linearly combine the different kernels instead. Amongst the studies that have applied this method are [30] (in that case solely refering to FDG-PET images) and [53] (where sMRI and DTI were combined).

Alternative approaches such as import vector machines (IVM) and regularized extreme learning machines (RELM) have also been used (in [42]). While in SVM the support vectors are used to build the model, in IVM these subsets of the input vectors are selected, using Kernel Logistic Regression (KLR), by minimizing the regularized cost function to reduce computation time. RELM consists of another effective solution which adopts reliability-based classification, thus combining the decrease in computational time provided by ELM and a slightly increased accuracy from the imposed sparsity condition.

Other strategies have also been used, such as k-Nearest Neighbors, which is intuitive and simple to apply and can also perform reasonably well on this computer aided diagnosis task, as shown in [12]. Classification algorithms such as linear discriminant analysis (LDA) and Gaussian discriminant analysis (GDA) have also been applied for this purpose (namely in [31] and [41]), the former of which consists of a specific case of the latter, assuming not only that the distribution of the data associated to each class follows a normal distribution, as the latter, but also that the data for all classes share the same covariance matrix.

Several studies have also opted for applying ensemble classifiers (namely using SVMs, once again, as in [48]), where multiple ”weak” classifiers are combined to build the final classification algorithm, as, by aggregating the predictions of multiple classifiers, this method may improve the generalization ability and robustness of the model and thus reduce possible overfitting problems. In this context, in [32], the authors used an ensemble of sparse representation-based classifiers (SRC), where each novel test sample is coded as a sparse linear combination of all training samples by l_1-norm minimization and then classified by evaluating which class produces the minimum reconstruction error. Another approach that has recently raised much interest consists of random forests (RF), again an ensemble learning machine, in this case composed of multiple binary decision trees, as applied in [30],[46]. Furthermore, extremely randomized trees (ET) have also been used for this purpose, namely in [30]. Similar to RF, this consists of a tree ensemble method, although in this case an additional step of randomization is performed, particularly regarding the split at each node, which contrary to other approaches is thus

randomly selected instead of computed to yield the optimal choice.

Regarding the performance of the proposed classification algorithms, in terms of the binary classification task between Alzheimer's Disease and cognitively normal subjects using image registration, an accuracy as high as 95.7% has been attained (namely in [49], using an SVM and following a multi-modal approach, including FDG-PET and sMRI, applied to a dataset of nearly 50 subjects of each class), while concerning the sensitivity that value corresponds to 98.78% (in [41], using a GDA classifier on 131 AD bearing subjects and 187 healthy ones, using sMRI) and in terms of specificity that corresponds to 98.2% (again in [49]). Furthermore, the maximum values of 100% for these metrics have been reported in [36]. As for the classification between cognitively normal and MCI subjects, an accuracy of 92.36% was achieved (in [13], using a deep learning strategy, namely a convolutional neural network pre-trained with a sparse autoencoder, on a set of 755 sMRI brain scans of each class), while a sensitivity as high as 99.58% (in [48], using a multi-modal approach combining FDG-PET and sMRI, and an SVM ensemble classifier on a set of 101 healthy subjects and 199 diagnosed with MCI) and specifity of 90.40% (in [32], using a dataset of 229 brain scans of cognitively normal subjects and 225 of MCI subjects, applying an ensemble of SRCs followed by a final SVM classifier) have also been reported. With regard to the deep learning strategies here introduced, other classifiers such as softmax have also been used, so that the following section will thus go into more detail regarding this method. The studies presented in this chapter also refer to other binary and multi-class problems, and are summarized in 3.7, although only the two aforementioned classification tasks are considered for the purpose of this thesis.

In terms of the proposed CAD systems that did not apply the image registration step, or only did so in the training stage, the highest accuracy reported for the classification problem between Alzheimer's disease and cognitively normal subjects was, for the first case, 89.7% (in [44]) and, for the latter, 92.75% (in [43]). In [44], where the image registration step was fully disregarded, the values reported for the sensitivity and specificity were, respectively, 88.0% and 92.6%, while the highest reported values for the studies that used this processing step only at the training stage (in [59] and [43]) were, respectively, 93.48% and 93.50% (in [43]). In [44] the performance of the CAD system in the classification task between cognitively normal and MCI bearing subjects was also evaluated, reaching values of accuracy, sensitivity and specificity of, respectively, 74.0%, 86.6% and 92.6%.

3.6 Deep Learning

In recent years, deep learning strategies have also become well established in the field of computer aided diagnosis of Alzheimer's disease, with several approaches having been explored and allowing to achieve state of the art performances. Indeed, methods such as convolutional neural networks (CNN), autoencoders (AE, or, specifically, sparse autoencoders (SAE)), deep Boltzmann machines (DBM) and DenseNets have been used for this purpose.

A commonly used strategy (applied in [13] and [50]) consists of combining the two former, using CNNs pre-trained with a SAE. Equivalently, this corresponds to applying a SAE for feature extraction, allowing to find the, resulting, appropriate set of filters to be used in the convolutional layer of the CNN.

After this convolution, the integration of a pooling layer then allows for a downsampling procedure, reducing the number of parameters (such as the weights and bias terms) to be learned through training of that neural network, and also providing invariance to small translations in the input data, ensuring that the algorithm can still perform well and is robust to these. To conclude the CAD procedure through this method, a final, classification layer is then used as the output layer. In case of a binary classification task, this typically consists of a softmax function, for which the output is in the range from 0 to 1, having these correspond to the two distinct classes considered.

Alternatively, a simple sparse autoencoder or stacked sparse autoencoder has been applied for the purpose of CAD of AD (in [47]). Indeed, the SAE-learned feature representation is equivalent to a filter response from the input image when convolved with the filters that resulted from training the network, which correspond precisely to the weight terms entering the hidden layer. Furthermore, using a stacked sparse autoencoder (SSAE) strategy, one can increase the depth of this neural network, and, since each hidden layer can be seen as a higher level representation of the previous layer (altough more abstract so that it is usually nontrivial to define the exact meaning of each of those layers), a more profound feature representation can be learned. The obtained representation can thus be used either for an output classification layer, as discussed for the previous case, or as the input to any other classification algorithm (namely an SVM, amongst others), concluding the CAD procedure.

Another strategy that has been applied (in [43]) consists of a multi-instance convolutional neural network (MICNN) model, an end-to-end classification model that is capable of learning local-to-global representations layer by layer. Particularly, local representations can first be learned via multiple sub-CNN architectures embedded with a series of convolutional, pooling and fully connected layers which, despite sharing the same structure, have different network paramaters so as to learn specific features for each local area. The obtained local feature representations can then be concatenated and provided as input to additional fully-connected layers, which can capture the complex relationship among local regions and thus represent the brain structure at a global level, having the output of the final fully connected layer be, once again, fed into a soft-max classifier.

In [44], another method was used, in that case corresponding to a DenseNet. Several advantages have been reported concerning this method, in contrast to the traditional CNN architecture. Indeed, not only can these alleviate the vanishing-gradient problem, by including a direct connection from the low to high level layers, but also substantially reduce the number of network parameters, while the reuse of some features, especially at the lower levels. These advantages thus enable DenseNets to achieve better performances in some applications than CNNs.

As mentioned, another deep learning model that has been applied in this context consists of a deep Boltzmann machine (DBM, in [48]), which corresponds to an undirected graphical model structured by hierarchically stacking multiple restricted Boltzmann machines (RBM), and that allows to find feature representations in a probabilistic manner. Once again, this hierarchical nature enables to capture highly nonlinear and complicated patterns or statistics such as the relations among the input data so that, unlike other methods that considered hand-crafted features or outputs from the predefined functions, the role of determining feature representations is thus assigned to the DBM, finding them in an unsupervised way,

similar to that of the previous deep learning strategies such as the AE. The final classification procedure is thus, again, the only supervised learning problem in terms of training the model. In this case, this can be performed by using the discriminative version of DBM, with the label information at the top layer.

Moreover, similar to the autoencoder (as will be explained in detail in section 4.1.1.3.1), the symmetricity of the RBM model (the building block for DBM construction) in terms of its connectivity between visible and hidden layer also allows for a reconstruction of the input data from the hidden representations, so that it is indeed also considered as an AE. However, unlike the stacked autoencoder model, the characteristic undirected graphical model methodology of the DBM also brings the advantage of having the approximate inference procedure after the initial bottom-up pass incorporate top-down feedback, allowing it to use higher-level knowledge to resolve uncertainty about intermediate-level features and creating better data-dependent representations and statistics for learning, which can eventually lead the DBM model to outperform SSAE. Moreover, from the perspective of a multi-modal approach, the bidirectional information flow provided by DBM (in that case between the different modalities considered) also enables it to obtain a shared representation, thus fusing the complementary information without possible loss of correlation, as would occur in traditional methods that combine that information after extracting each modality-specific features.

3.7 Summary

The following tables summarize the considerations presented throughout this section, regarding neuroimaging modalities, feature extraction and selection techniques as well as feature transformation methods, classification algorithms and the resulting state of the art performances in terms of accuracy, sensitivity and specificity. In contrast to the previous separation between deep learning methods and supervised learning problems, in these tables the two are grouped together in terms of the objective for which they are used, according to the referred list.

To account for the fact that the model performance comparison throughout these studies might not be fair due to possible differences in the quality and amount of available data for training and testing the proposed CAD systems, information regarding the experimental setup, in terms of the number of participants that constitute the dataset used in each case, is also provided.

Table 3.1: Summary of the state of the art on CAD of Alzheimer's Disease. Acronyms: Accuracy (ACC); Sensitivity (SEN); Specificity (SPE). Codes for the different classification tasks: 1 (CN/AD), 2 (CN/MCI), 3 (AD/MCI), 4 (MCIc/sMCI), 5 (CN/MCIc), 6 (CN/sMCI), 7 (AD/sMCI), 8 (AD/MCIc), 9 (CN/AD/MCI), 10 (CN/ AD and MCI), 0 (CN/AD/MCIc/sMCI).

Author(s)	Image Registration	Imaging Technique(s)	Feature Extraction	Feature Transformation	Feature Selection	Classification	Nº of Participants	ACC|SEN|SPE (%)
Morgado et. al, 2013 [12]	Yes	FDG-PET	Voxel-based; ROI	VI; Textons	MI; Wrapper	k-NN; SVM	59 AD; 70 CN; 104 MCI	91.4|– |– (1) 74.9|– |– (2)
Garali et. al, 2015 [29]	Yes	FDG-PET	ROI	4 moments of VI+ROI entropy	SPF; RF	SVM; RF	81 AD; 61 CN	95.07|– |– (1)
Wehenkel et. al, 2018 [30]	Yes	FDG-PET	ROI	VI	RF; ET; MKL	RF; ET; MKL	22 MCIc; 23 sMCI	80.44|78.18|82.61 (4)
Ramírez et. al, 2013 [45]	Yes	SPECT	ROI	VI; SD and inter-slice corr.	FDR (1st,2nd order stat.)	SVM	29 AD; 23 CN	84.62|– |– (1); 90.38|93.10|86.96 (1)
Coupé et. al, 2012 [31]	Yes	sMRI	Patches + ROI	SNIPE HC and EC grading and volume	—	LDA	198 AD; 231 CN; 167 MCIc; 238 sMCI	89|84|93 (1); 71|70|72 (4); 86|80|89 (5); 69|76|63 (6); 77|77|78 (7); 62|63|60 (8)
Liu et. al, 2012 [32]	Yes	sMRI	Patches	GM density	Downsamp. + t-test	SRCs ens.; SVM	198 AD; 229 CN; 225 MCI	90.80|86.32|94.76 (1) 87.85|85.26|90.40 (2)
Liu et. al, 2014 [33]	Yes	sMRI	Patches	GM density + Mean + SD + Inter-patch corr.	Downsamp. + t-test	SVM	198 AD; 229 CN; 225 MCI	92.0|90.9|93.0 (1) 85.3|82.3|88.2 (2)

Table 3.1 Continued from the previous page.

Author(s)	Image Registration	Imaging Technique(s)	Feature Extraction	Feature Transformation	Feature Selection	Classification	Nº of Participants	ACC|SEN|SPE (%)
Khedher et. al, 2015 [34]	Yes	sMRI	Voxel-based	VI	VI Threshold + PLS; PCA	SVM	188 AD; 229 CN; 401 MCI	88.49|85.11|91.27 (1) 81.89|82.16|81.62 (2) 87.03|88.65|85.41 (3)
Zhang et. al, 2015 [35]	Yes	sMRI	Whole brain + ROI	MIE Eigenvalues	ICV + PCA + t-test	SVM	28 AD; 98 CN	92.36|83.48|94.90 (1)
Payan & G. Montana, 2015 [13]	Yes	sMRI	Whole brain	CNN Feat. Maps (VI - Trained w/ SAE)	Max-pooling	Softmax	755 AD; 755 CN; 755 MCI	95.39|– |– (1) 92.11|– |– (2) 86.84|– |– (3) 89.47|– |– (9)
Dessouky et.al, 2016 [36]	Yes	sMRI	Whole brain	DTT	Image Reduction	SVM	49 AD; 71 CN	100|100|100 (1)
Jongkreangkrai et. al, 2016 [37]	Yes	sMRI	ROI	HC and Amygdala Rel. Volumes + EC thickness	—	SVM	100 AD; 100 CN	—
Zhang et. al, 2016 [38]	No	sMRI	Patches + Voxel-based	LEP + 3D displacements	t-test + Variance + Entropy + RF	SVM	199 AD; 229 CN	83.7|80.9|86.7 (1)
Beheshti et. al, 2017 [39]	Yes	sMRI	ROI+Voxel-based	VI	VBM+t-test + GA	SVM	92 AD; 94 CN; 71 MCIc; 65 sMCI	93.01|89.13|96.80 (1) 75.00|76.92|73.23 (4)
Khedher et. al, 2017 [40]	Yes	sMRI	Whole brain	ICAm subspace projection	ICAm	SVM	188 AD; 229 CN; 398 MCI	89.5|92.4|86.6 (1) 79.6|82.9|76.5 (2) 86.4|86.6|86.1 (3)

Table 3.1 Continued from the previous page.

Author(s)	Image Registration	Imaging Technique(s)	Feature Extraction	Feature Transformation	Feature Selection	Classification	N^o of Participants	ACC|SEN|SPE (%)
Fang et. al, 2017 [41]	Yes	sMRI	ROI	7 ROI Shape Measures	ANOVA + Inc. Error	GDA	131 AD; 187 CN; 301 MCI	93.28|98.78|81.08 (10) 81.71|71.43|85.25 (3)
Lama et. al, 2017 [42]	Yes	sMRI	ROI	5 ROI Shape Measures	PCA	SVM; IVM; RELM	70 AD; 70 CN; 74 MCI	80.32|87.10|90.63 (1) 61.58|60.78|66.89 (9)
Liu et. al, 2018 [43]	No	sMRI	Voxel-based + Patches	MICNN Feature Maps (VI)	t-test + RF + Max-pooling	Softmax	199 AD; 229 CN; 167 MCIc; 226 sMCI (ADNI 1) 159 AD; 200 CN; 38 MCIc; 239 sMCI (ADNI2) 46 AD; 23 CN (MIRIAD)	91.09|88.05|93.50 (1; ADNI2) 92.75|93.48|91.30 (1; MIRIAD) 76.90|42.11|82.43 (4; ADNI2)
Li & Liu, 2018 [44]	No	sMRI	Patches	DenseNet Feature Maps (VI)	PCA + Max-pooling	Softmax	199 AD; 229 CN; 403 MCI	89.7|88.0|92.6 (1) 74.0|86.6|57.2 (2)
Cui et. al, 2011 [52]	Yes	sMRI + CSF + Neur. Tests	ROI	ROI TA+CV+SV + T-tau/$A\beta 42$ + P-tau/$A\beta 42$ + FAQ + LM + AVLT	mRMR + Wrapper	SVM	96 AD; 111 CN; 56 MCIc; 87 sMCI	67.13|96.43|65.52 (4)
Liu et. al, 2014 [47]	Yes	sMRI+FDG-PET	Whole brain	SAE Feature Maps (GM vol. + FDG-PET VI)	Sparsity Constraint	Softmax	65 AD; 77 CN; 67 MCIc; 102 sMCI	87.76|88.57|87.22 (1) 76.92|74.29|78.13 (2) 47.42|65.71|83.75 (0)

Table 3.1 Continued from the previous page.

Author(s)	Image Registration	Imaging Technique(s)	Feature Extraction	Feature Transformation	Feature Selection	Classification	N^o of Participants	ACC|SEN|SPE (%)
Suk et. al, 2014 [48]	Yes	sMRI + FDG-PET	Voxel-based + Patches	DBM Feature Maps (GM dens. + FDG-PET VI)	t-test	SVM Ensemble	93 AD; 101 CN; 71 MCIc; 128 sMCI	95.35|94.65 |96.33 (1) 85.67|99.58|65.87 (2) 75.92|48.04|96.55 (4)
Zhu et. al, 2017 [49]	Yes	sMRI + FDG-PET	ROI	ROI Avg. VI + GM volume	Relational Reg.	SVM	51 AD; 52 CN 43 MCIc; 56 sMCI	95.7|96.6|98.2 (1) 79.9|97.0|59.2 (2) 72.4|49.1|94.6 (4)
Vu et. al, 2017 [50]	Yes	sMRI + FDG-PET	Patches + Whole brain	CNN Feat. Maps (VI, Trained w/ SAE)	Max-pooling	Softmax	145 AD; 172 CN	91.1|– |– (1) 89.2|– |– (2)
Mikhno et. al, 2012 [51]	Yes	sMRI + FDG-PET + PIB-PET	ROI	HC 3D ZD + HC Volume + ROI Avg. VI	Logistic Regression AUC	SVM	17 AD; 17 CN; 22 MCI	90.9|91.1|93.8 (1) 76.2|76.1|78.1 (2) 84.3|80.9|87.3 (3)
Tong et. al, 2017 [46]	Yes	sMRI + FDG-PET + CSF + *APOE*	ROI + Whole brain	NGF of volume + VI + $A\beta 42$ + T-tau + P-tau + *APOE*	RF	RF	37 AD; 35 CN; 75 MCI	91.8|88.9|94.7 (1) 79.5|85.1|67.1 (2) 60.2|– |– (9)
Ahmed et. al, 2017 [53]	Yes	sMRI + DTI MD	ROI + Patches	Hist. of GL-CHFs + CSF volume	–	MK-SVM	45 AD; 52 CN; 58 MCI	90.20|82.92|97.20 (1) 79.42|71.58|86.05 (2) 76.63|65.62|81.33 (3)
Bron et. al, 2017 [54]	Yes	sMRI + DTI + ASL	Whole brain	GM vol. + WM vol. + FA WM VI + CBF GM VI	–	SVM	24 AD; 34 CN	90|– |– (1)

Chapter 4

Materials and Methods

4.1 Theoretical and Mathematical Framework

4.1.1 Feature extraction and transformation

According to what was mentioned in the previous sections, several approaches, instead of one single method, were explored for the purpose of this work. Regarding feature transformation, both voxel intensity, histogram of textons and SSAE feature representations were considered.

4.1.1.1 Voxel intensity

The simplest choice of features used for image classification consists of the raw voxel intensity. Indeed, these are obtained directly from the FDG-PET brain scans, precisely corresponding to the image intensity at each voxel, and thus constituting a direct measure of the FDG uptake detected in each voxel.

4.1.1.2 Histogram of textons

As introduced in section 3.4, in this bag-of-features approach, the texture information of the images to be classified is considered. In particular, a texture is characterized by its responses to a set of orientation and spatial-frequency selective linear filters, which constitutes the filter bank [61]. Since by definition textures have spatially repeating properties, the resulting filter responses shouldn't be totally different at each pixel over the texture; on the contrary, while there should be several distinct filter response vectors, the remaining ones should simply be noisy variations of them, so that the filter responses can be clustered into a small set of prototype response vectors [61]. The resulting exemplar filter responses in turn correspond to the designation of textons. After using a set of training images to build the texton dictionary, model extraction is performed so that each image is represented by a histogram of texton frequencies [62]. The same model extraction procedure is performed on the test set so that the resulting histogram can be given as input to the classifier for the desired class label to be obtained [62].

This classification algorithm comprises, thus, a learning and a classification stage in a series of sequential steps, as will be explained in more detail in the sections that follow.

4.1.1.2.1 Filter responses

In order to obtain the filter responses that will characterize each texture, a filter bank is used. Amongst the filter sets commonly used for this texton-based approach are the ones proposed by Leung and Malik (LM), Schmid, and the Maximum Response sets [62]. Unlike the former, the filters in the Schmid and Maximum Response sets are rotationally invariant [62]. This is an interesting property for the purpose of this work, as for one particular dataset used, the images were allowed to appear at any orientation and the proposed methods should still be capable of overcoming the inherent classification challenge.

Regarding the LM set, a total of 48 filters are included, namely edge, bar and spot filters at multiple scales and orientations (particularly 2 Gaussian derivative filters at 6 orientations and 3 scales, 8 Laplacian of Gaussian filters and 4 Gaussian filters) [61]. The Schmid set, on the other hand, comprises 13 isotropic and rotationally invariant filters, of the form $F(r,\sigma,\tau) = F_0(\sigma,\tau) + cos(\frac{\pi\tau r}{\sigma})e^{\frac{-r^2}{2\sigma^2}}$ [63]. To enable feature extraction from anisotropic textures, which would not possible using solely the isotropic filters in the Schmid set, the Maximum Response set was proposed in [62], particularly the MR8 set, shown in Figure 4.1. This includes a total of 38 isotropic and anisotropic filters at multiple scales and orientations (namely Gaussian and Laplacian of Gaussian at a single scale as the two isotropic filters, as well as an edge filter and a bar filter at 6 orientations and 3 scales each). Nonetheless, although this results in 38 filter responses, only the maximum across all orientations (regarding the edge and bar filters) is kept, which achieves the desired rotation invariance as well as reduces the dimensionality of the filter response space where textons are searched for, particularly from 38 to 8 filter responses.

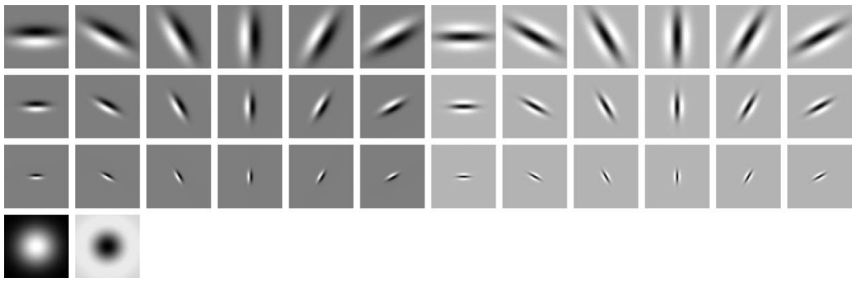


Figure 4.1: Representation of the MR8 filter bank.

A 3D version of the MR8 filter bank has also been proposed in [12]. More concretely, the proposed filter bank was composed by a 3D Gaussian filter and its Laplacian, and three other types of filters, namely, 3D edge filters, bar and plane filters (at 3 triplets of scales and 61 orientations). The mathematical formulation for these filters (or an example of which, for the three anisotropic filters, in this case with rotational symmetry around the x_3 axis) is presented in equations 4.1 to 4.5, respectively:

$$G(x,\sigma) = K\ exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{3}x_i^2\right\} \tag{4.1}$$

$$\nabla^2 G(.) = -\frac{K}{\sigma}\left(1 - \frac{\sum_{i=1}^{3}x_i^2}{\sigma^2}\right)\ exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{3}x_i^2\right\} \tag{4.2}$$

$$\frac{\partial G(.)}{\partial x_3} = -\frac{K\ x_3}{\sigma_3^2}\ exp\left\{-\frac{1}{2}\sum_{i=1}^{3}\frac{x_i^2}{\sigma_i^2}\right\} \tag{4.3}$$

$$\frac{\partial^2 G(.)}{\partial x_3^2} = -\frac{K}{\sigma_3^2}\left(1-\frac{x_3^2}{\sigma_3^2}\right)\ exp\left\{-\frac{1}{2}\sum_{i=1}^{3}\frac{x_i^2}{\sigma_i^2}\right\} \tag{4.4}$$

$$\sum_{i=1}^{2}\frac{\partial^2 G(.)}{\partial x_i^2} = -K\sum_{i=1}^{2}\left(\frac{\sigma_i^2-x_i^2}{\sigma_i^4}\right)\ exp\left\{-\frac{1}{2}\sum_{i=1}^{3}\frac{x_i^2}{\sigma_i^2}\right\} \tag{4.5}$$

where K is a normalization factor, $x = (x_1, x_2, x_3)$ is the vector of spatial coordinates and $(\sigma_1, \sigma_2, \sigma_3)$ corresponds to the scale triplet. An example of each of these filters is depicted in Figure 4.2, retrieved from [12].

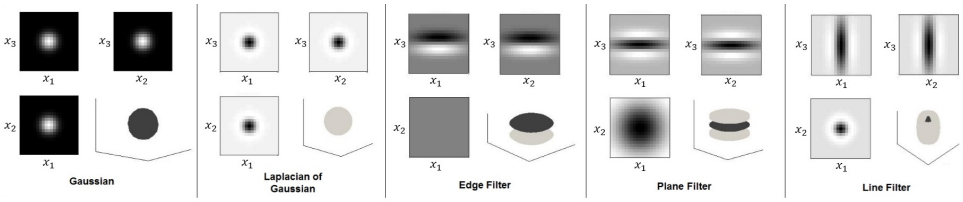


Figure 4.2: Representation of an example of each type of filter in the 3D extension of the MR8 set.

The convolution of each texture image with this filter bank thus results in a filter response space where each voxel is transformed to a N_{filt}-dimensional vector (where N_{filt} is the number of filters used, in this case, 11), having each dimension represent the response to the corresponding filter. Moreover, according to the original approach as proposed in [62], three pre-processing steps should also be performed prior to the learning stage. In particular, the input images should be normalized in its intensity to have zero mean and unit variance, each filter should be normalized to have unit l_1 norm, as given by:

$$\|v\|_1 = \sum_{k=1}^{n}|v_k| \tag{4.6}$$

for a generic n-dimensional vector v, and the filter response at each voxel position x, $F(x)$, should also be normalized according to the following equation (motivated by Webber's law) [62]:

$$F(x) \leftarrow \frac{F(x)}{\|F(x)\|_2}\ log\left(1+\frac{\|F(x)\|_2}{0.03}\right) \tag{4.7}$$

where $\|F(x)\|_2$ represents the l_2 norm of $F(x)$, which, for the same generic vector v, is given by:

$$\|v\|_2 = \sqrt{\sum_{k=1}^{n}|v_k|^2} \tag{4.8}$$

4.1.1.2.2 Building the dictionary

Having obtained the filter responses for each pixel (or in the 3D case, for each voxel), a number of training images of each class is used to build the texton dictionary, as a part of the learning stage of the algorithm. Depending on the application, it may not be required that all training images are used at this

point; instead, in some cases, a smaller amount can be (for instance, randomly) picked for that purpose, as long as possible relevant features occurring in the remainder aren't discarded, which could potentially induce misclassification errors [12]. For each class, all the filter response vectors for all the used training images are then clustered using an algorithm such as K-means.

As an unsupervised learning problem, some assumptions are required for data clustering with k-means. Particularly, it is assumed that each cluster is approximately isotropic and well represented by a prototype (the centroid), $c_j \in \mathbb{R}^p$, that each data point is closer to the nearest point (to which the distance should ideally be small) in the same cluster than to all the ones in the remaining clusters and that each of them has a uniform density of data. This corresponds to a global optimization problem where the cost function to be minimized is given by [64]:

$$C = \sum_{i=1}^{K} \sum_{x \in T_i} \|x - c_i\|^2 \tag{4.9}$$

where $T = \{x_1, ..., x_n\}$ corresponds to the feature vectors to be clustered (the training set), T_i to the ones assigned to the i-th cluster, and K designates the total number of clusters. As this joint optimization with respect to the assignment of each feature vector to a cluster becomes unfeasible, a separate optimization can be performed, one regarding data assignment and the other centroid update [64]. This is, indeed, iteratively performed in K-means, for which the criterion is to find K centroids (corresponding to the center of the clusters, which in the current work correspond to the textons) such that after assigning each vector to the nearest center, the sum of squared distances from the centers are minimized [64],[61]. Following centroid initializaion (namely by attributing them to a set of randomly picked feature vectors), the two operations are thus repeated until this partitional algorithm converges and a local minimum of the cost function is achieved [64]. The mathematical formulation for data assignment and centroid update as performed by K-means is presented in equations 4.10 and 4.11, respectively:

$$x \in T_j : j = arg\ \min_i \|x - c_i\|^2 \tag{4.10}$$

$$c_j = \frac{1}{\#T_j} \sum_{x \in T_j} x \tag{4.11}$$

An alternative clustering algorithm that can be used for this same purpose is sequential K-means (or online K-means). In this iterative method, applied until interrupted, data assignment and centroid update are performed sequentially, as the name indicates. Thus, instead of assigning all the training data to each corresponding cluster and then updating its centroids, the procedure is performed on each training sample at a time. This means that each training sample is picked (namely after randomly ordering the training data) and the two steps are applied, so that it is assigned a cluster and the corresponding centroid is immediately updated, prior to considering the following training sample.

The resulting textons from the different texture classes, which correspond precisely to the obtained cluster centroids, are then combined to form the texton dictionary. This step of the algorithm's learning stage is illustrated in Figure 4.3, adapted from [62]. In the current work, the texture class corresponds to

one of three possibilities, namely cognitively normal, diagnosed with MCI or diagnosed with AD.

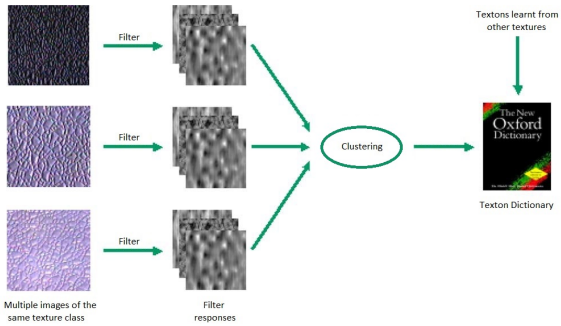


Figure 4.3: Illustration of the procedure for building the texton dictionary.

4.1.1.2.3 Extracting the models

Having built the dictionary of textons, each filter response and hence each image pixel (or voxel, in the 3D case) is assigned one of the texton labels from those within the dictionary, specifically the one for which the distance to the filter response vector being considered is the minimum, which is equivalent to using a nearest neighbor classifier.

Considering the training set $T = \{(x_i, y_i), i = 1, ..., n\}$, with y corresponding to the class label and $x \in \mathbb{R}^p$ corresponding to the input vector, the strategy inherent to the nearest neighbor classifier method to predict y for a new x vector would consist of finding the training pattern x_i nearest to x (using the Euclidean distance) and approximating y by y_i [65]. Considering $\{(x_{(1)}, y_{(1)}), ..., (x_{(n)}, y_{(n)})\}$ as a re-ordering of the training set such that

$$\|x_{(1)} - x\| \leq \|x_{(2)} - x\| \leq ... \leq \|x_{(n)} - x\|$$

this method would thus assign x to the outcome of the nearest neighbor: $f(x) = y_{(1)}$. Adapting this formulation to the problem at hand, each training pattern x_i would correspond to a texton (the N_{filt} dimensional vector) and y_i to its label, while the new vector x would be the filter response to be assigned a texton, in that case y. The model for each training image, as given by the histogram of textons, can then be generated, and represent it from that point forward, which concludes the learning stage of the general classification algorithm. Figure 4.3, once again adapted from [62], depicts this model extraction step for an example training image, following its convolution with the filter bank and mapping of the resulting filter responses to the textons used to build the dictionary.

It should be highlighted that, while figures 4.3 and 4.4 refer to the two-dimensional case, the training and test images used in the current work are in fact three-dimensional, as mentioned.

In what refers to the classification stage, each novel image is mapped to a texton distribution through the same model extraction procedure and the resulting histogram is classified, as will be presented in section 4.1.2, on the basis of comparison with the models learned during training [62]. It is important to note that all the referred histograms should be normalized to sum to unity, although this isn't required if the number of features considered for the classification (and thus the total number of textons) is the same across all images.

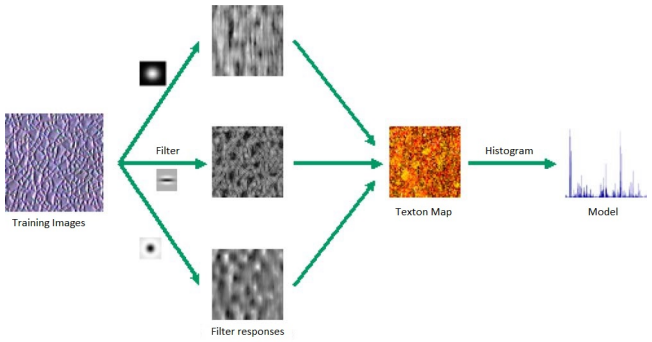


Figure 4.4: Illustration of the model extraction procedure.

4.1.1.3 SSAE feature representations

As mentioned in the previous chapters, a deep learning strategy was also explored and compared against the remaining proposed approaches, particularly a Stacked sparse autoencoder (SSAE). As the name indicates, this consists of multiple layers of basic sparse autoencoders, in which the outputs of each layer are wired to the inputs of the successive layer. A sparse autoencoder (SAE) derives, in turn, from enforcing a sparsity constraint on its simplest version, the autoencoder, which consists of an unsupervised feature learning algorithm that can be used to obtain a feature representation of high dimensional input data by finding some correlation among it [66]. This results in a set of low-dimensional, high-level features [14], thus performing feature transformation, in a similar way to what is attained by applying the set of pre-defined filters in the texton-based approach as presented in 4.1.1.2.

4.1.1.3.1 The autoencoder

In detail, an autoencoder is a multi-layer feed-forward neural network, thus including the input layer, a hidden layer and the output layer, without any directed loops or cycles [67], as presented in Figure 4.5 (adapted from [14]). It aims at minimizing the discrepancy between input and reconstruction by learning an encoder and a decoder, yielding a set of weights W and biases b [66]. Defining the unlabeled training data as $\{x^{(1)}, ..., x^{(n)}\}$, n being the number of training samples, and the target values of this neural network as $\{y^{(1)}, ..., y^{(n)}\}$, $x^{(i)} \in \mathbb{R}^p, y^{(i)} \in \mathbb{R}^p, \forall i$, an autoencoder thus aims at setting the target values to be equal to the inputs, using $y^{(i)} = x^{(i)}, \forall i$ [67]. This is equivalent to saying that this algorithm tries to learn an approximation to the identity function, so as to output a reconstructed $\hat{x}$ that is similar to x, where this discrepancy to be minimized is described by an average sum-of-squares error term in the cost function, as given by:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\left\|h_{W,b}(x^{(i)}) - y^{(i)}\right\|^2\right) \tag{4.12}$$

where $h_{W,b}(x^{(i)})$ corresponds to the result of the output layer, for each input training pattern $x^{(i)}$, thus equivalent to its output reconstruction [67].

If there is structure in the data (if it isn't completely random), by limiting the number of hidden units or enforcing a sparsity constraint (in that case, possibly still allowing for a large number of hidden units), the network can then detect it; particularly, if the number of hidden units is much smaller than the input and

output layer, the network is forced to learn a compressed representation of the input, thus allowing to identify existing correlations in the input features, as mentioned [67]. As for the sparsity constraint, this consists of enforcing the neurons to be inactive most of the time, which would correspond to an output value close to 0 (in contrast to an output value close to 1 if the neuron is active) [67].

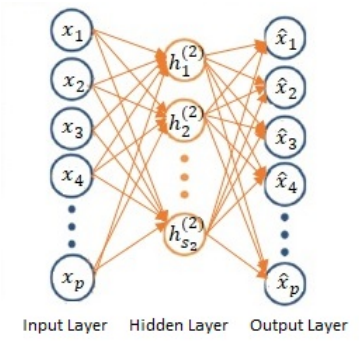


Figure 4.5: Representation of the architecture of a typical autoencoder. For simplicity, the bias (intercept) terms for the input and hidden layers are not shown.

Considering the general autoencoder architecture, assuming $W_{ij}^{(l)}$ denotes the weight associated with the connection between unit j in layer l and unit i in layer $l+1$ and that $b_i^{(l)}$ is the bias associated with unit i in layer $l+1$, the total weighted sum of inputs to unit i in layer $l+1$, $z_i^{(l+1)}$, is given by:

$$z_i^{l+1} = \sum_{j=1}^{s_l} W_{ij}^{(l)} a_j^{(l)} + b_i^{(l)} \tag{4.13}$$

where s_l corresponds to the number of units in layer l (particularly being equal to p in the input layer, $l=1$), and $a_j^{(l)}$ to the activation of unit j in layer l, equivalent to writing $a_i^{(l)} = f(z_i^{(l)})$ (again, particularly for the input layer, we have that $a_j^{(1)} = x_j$). In turn, f is the activation function, particularly chosen to be the logistic sigmoid function (for which the output is in the range $[0,1]$), as given by:

$$f(z) = \frac{1}{1+e^{-z}} \tag{4.14}$$

To help prevent overfitting, an extra, regularization, term is also usually introduced in the cost function, complementing the formulation in 4.12. Particularly, this consists of a weight decay term, which tends to decrease the magnitude of the weights, as given by:

$$\frac{\lambda}{2} \|W\|_2^2 = \frac{\lambda}{2} \left(\sum_{l=1}^{n_l} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} W_{ji}^{(l)} \right) \tag{4.15}$$

where λ is the weight decay parameter [66]. The overall cost function is thus given by:

$$J(W,b) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} \|h_{W,b}(x) - y\|^2 \right) + \frac{\lambda}{2} \left(\sum_{l=1}^{n_l} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} W_{ji}^{(l)} \right) \tag{4.16}$$

In order to train the SAE network to obtain the set of weights that minimize the cost function, a commonly used optimization algorithm consists of the gradient descent. This results in an update of the

weights and bias terms as given by:

$$W_{ij}^{(l)} = W_{ij}^{(l)} - \alpha \frac{\partial}{\partial W_{ij}^{(l)}} J(W,b) \tag{4.17}$$

$$b_{i}^{(l)} = b_{i}^{(l)} - \alpha \frac{\partial}{\partial b_{i}^{(l)}} J(W,b) \tag{4.18}$$

where α corresponds to the specified learning rate (or step size). The partial derivates $\frac{\partial J(W,b)}{\partial W_{ij}^{(l)}}$ and $\frac{\partial J(W,b)}{\partial b_{i}^{(l)}}$ are then computed with resource to the backpropagation algorithm, as is the case in general neural networks, applied after each forward propagation.

The first step of the training stage thus consists of the initialization of the weights and bias terms, $W_{ji}^{(l)}$ and $b_{i}^{(l)}$, prior to performing the first forward pass. Many initialization techniques have been proposed which generally perform well, one of which consists of the choice of small random values, close to 0, namely sampling from a zero-mean Gaussian distribution with a very small standard deviation [68]. Particularly for the weights initialization, the randomness should be introduced to prevent all the hidden units to learn the same function of the input features. As for the choice of very small values, close to 0, this is also highly relevant since backpropagation requires that the gradient of the activation function is different to 0, and, recalling that the activation function being used here is the logistic function, which saturates for very large positive values and very small negative values, by choosing these weights close to 0 we enforce that those gradient values differ from 0 and thus that the network will be able to learn.

After weight initialization, a forward pass is performed on each training example, and the weighted sum of inputs and corresponding activation for each unit is obtained, including the output value $h_{W,b}(x)$. The weights and biases are then updated according to 4.17, by applying the backpropagation algorithm. For each training pattern x, this is based on the chain rule for differentiation, holding, for the output layer:

$$\frac{\partial J(W,b;x,y)}{\partial W_{ij}^{(n_l-1)}} = \frac{\partial J(W,b;x,y)}{\partial z_{i}^{(n_l)}} \frac{\partial z_{i}^{(n_l)}}{\partial W_{ij}^{(n_l-1)}} \tag{4.19}$$

which is equivalent to writing:

$$\frac{\partial J(W,b;x,y)}{\partial W_{ij}^{(n_l-1)}} = \epsilon_{i}^{(n_l)} a_{j}^{(n_l-1)} \tag{4.20}$$

having:

$$\epsilon_{i}^{(n_l)} = \frac{\partial J(W,b;x,y)}{\partial z_{i}^{(n_l)}} = \frac{\partial J(W,b;x,y)}{\partial a_{i}^{(n_l)}} \frac{\partial a_{i}^{(n_l)}}{\partial z_{i}^{(n_l)}} = \frac{\partial J(W,b;x,y)}{\partial a_{i}^{(n_l)}} \, f'(z_{i}^{(n_l)}) \tag{4.21}$$

where $J(W,b;x,y)$ corresponds precisely to the cost function for each training pattern x. Considering the general definition in 4.29, for these output nodes we can directly measure the difference between the network's activation and the true target value, obtaining:

$$\epsilon_{i}^{(n_l)} = \frac{\partial}{\partial z_{i}^{(n_l)}} \frac{1}{2} \left\| y - h_{W,b}(x) \right\| = -(y_i - a_{i}^{(n_l)}) \cdot f'(z_{i}^{(n_l)}) \tag{4.22}$$

As for the hidden units, the computation must take into account how each node was responsible for any errors in the following layers up to the output one. It is hence based on a weighted average of the

error terms of the nodes that use $a_i^{(l)}$ as an input, as derived from:

$$\epsilon_i^{(l)} = \frac{\partial J(W,b;x,y)}{\partial z_i^{(l)}} = \sum_{j=1}^{s_{l+1}} \left(\frac{\partial J(W,b;x,y)}{\partial z_j^{(l+1)}} \right) \left(\frac{\partial z_j^{(l+1)}}{\partial a_i^{(l)}} \right) \left(\frac{\partial a_i}{\partial z_i^{(l)}} \right) = \left(\sum_{j=1}^{s_{l+1}} W_{ji}^{(l)} \epsilon_j^{(l+1)} \right) f'(z_i^{(l)}) \quad (4.23)$$

Once again the expression in 4.20, still holds, so that the partial derivatives used in expressions 4.17 and 4.18 are, in turn, given by:

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W,b;x,y) = a_j^{(l)} \epsilon_i^{(l+1)} \quad (4.24)$$

$$\frac{\partial}{\partial b_i^{(l)}} J(W,b;x,y) = \epsilon_i^{(l+1)} \quad (4.25)$$

4.1.1.3.2 Sparse autoencoders

As mentioned, sparse autoencoders consist of an extension to the autoencoder architecture, by imposing a further sparsity constraint. In fact, denoting $a_j^{(2)}(x^i)$ as the activation of hidden unit j when the network is given a specific input $x^{(i)}$, and $\hat{\rho}_j$ as the average activation of hidden unit j, given by:

$$\hat{\rho}_j = \frac{1}{n} \sum_{i=1}^{n} \left[a_j^{(2)}(x^i) \right] \quad (4.26)$$

enforcing the sparsity constraint is equivalent to enforcing that $\hat{\rho}_j = \rho$, ρ being the sparsity parameter, typically close to 0, so that the majority of the hidden units' activations must be near 0 [67]. This implies that an extra penalty term (besides the average sum-of-squares error term) should be added to the overall cost function to penalize $\hat{\rho}_j$ deviating significantly from ρ, typically based on the Kullback-Leibler (KL) divergence, a standard function for measuring how different two distributions are [67]. This penalty term is thus given by:

$$\beta \sum_{j=1}^{s_2} KL(\rho \parallel \hat{\rho}_j) \quad (4.27)$$

where β controls the weight of the sparsity penalty term, s_2 is the number of units in the hidden layer and $KL(\rho \parallel \hat{\rho}_j)$ is the is the KL divergence between a Bernoulli random variable with mean ρ and a Bernoulli random variable with mean $\hat{\rho}_j)$, given by:

$$KL(\rho \parallel \hat{\rho}_j) = \rho \, log \frac{\rho}{\hat{\rho}_j} + (1-\rho) log \frac{1-\rho}{1-\hat{\rho}_j} \quad (4.28)$$

which is a convex function that reaches its minimum of 0 at $\rho = \hat{\rho}_j$ [67]. This results, in turn, in an overall cost function given by:

$$J(W,b) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} \left\| h_{W,b}(x) - y \right\|^2 \right) + \frac{\lambda}{2} \left(\sum_{l=1}^{n_l} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} W_{ji}^{(l)} \right) + \beta \sum_{j=1}^{s_2} KL(\rho \parallel \hat{\rho}_j) \quad (4.29)$$

The same procedure for weight and bias terms' update is followed, although with minor modifications to account for the KL divergence term being introduced. Specifically, applying the backpropagation

algorithm to compute $\epsilon_i^{(2)}$ (the term relative to the hidden layer units), is no longer performed using expression 4.30, but instead using [67]:

$$\epsilon_i^{(l)} = \left(\sum_{j=1}^{s_{l+1}} W_{ji}^{(l)} \epsilon_j^{(l+1)} + \beta \left(\frac{-\rho}{\hat{\rho}_i} + \frac{1-\rho}{1-\hat{\rho}_i} \right) \right) f'(z_i^{(l)}) \tag{4.30}$$

Furthermore, since the average activation $\hat{\rho}_i$ is required for this computation, the forward propagation of all training samples must be performed prior to the backpropagation on any of those samples.

4.1.1.3.3 Stacked sparse autoencoders

As for the stacked sparse autoencoder, as mentioned, it consists of multiple layers of basic shallow sparse autoencoders, in which the outputs of each layer are wired to the inputs of the successive layer. An example of an SSAE which consists of two basic shallow sparse autoencoders is shown in Figure 4.6, adapted from [14].

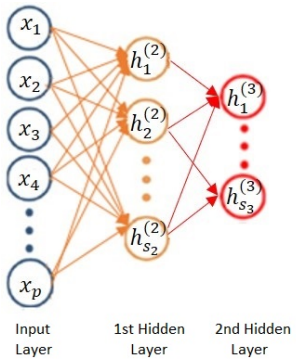


Figure 4.6: Representation of the architecture of a stacked sparse autoencoder (with two hidden layers). For simplicity, the decoder parts of each basic SAE in the figure are not shown.

Moreover, similar to the basic SAE, training an SSAE involves finding the optimal parameters (W, b) that minimize the discrepancy between the input and its reconstruction. After the optimal parameters are obtained, the SSAE yields a function $f : \mathbb{R}^p \rightarrow \mathbb{R}^{d_{h^{(3)}}}$ (where $d_{h^{(3)}}$ corresponds to the number of units in the third layer, and second hidden layer, which is thus equivalent to writing $f : \mathbb{R}^p \rightarrow \mathbb{R}^{s_3}$ according to the previous notation) that transforms the input feature vector to a new feature representation $h^{(3)} = f(x)$ [14]. Particularly for the case of an image patch, as performed in this work, this function thus transforms the input voxel intensities of each patch to its new feature representation to be further classified, as detailed in section 4.2.3.

Considering that there are multiple layers in this architecture, in order to obtain those optimal parameters, a greedy layer-wise training approach can be used, where the network layers are trained one at a time, so that the network is first trained having only one hidden layer, and only then it is trained with the two of those [67]. This stacked network, particularly with two hidden layers (as applied in this work), can then be constructed following three steps. First, the shallow sparse autoencoder on the raw inputs x to learn the primary feature representation $h^{(2)}$ is trained. Then, this primary feature representation is used as raw input to another shallow sparse autoencoder to learn the secondary feature activations $h^{(3)}$.

Finally, the two SAE are stacked together to form an SSAE with two hidden layers, which transforms an input $x \in \mathbb{R}^p$ to a deep feature representation $h^{(3)} = f(x)$ [69]. All the training samples can thus be written as $\left\{h^{(3)}(k)\ ,\ y(k)\right\}, k \in \{1, ..., n\}$, n being the number of training samples, and this denoting its respective pair of high-level features and label (having $y(k) \in \{0, 1\}, \forall k$, for the classification task at hands). It is important to highlight that, as mentioned, in the SSAE learning procedure, the label information isn't used at any point, hence consisting, in fact, of an unsupervised learning scheme [14].

In order to classify the obtained secondary features, following the high-level feature learning procedure enabled by the SSAE, an extra (output) layer can then be introduced, particularly corresponding to a softmax classifier as presented in the following section [14], [67].

Although with this approach the primary and secondary feature representations also correspond to filter responses (in a hierarchical feature learning procedure, where the second set of filters, $W^{(2)}$, are able to detect more complex and in-depth features than the former, $W^{(1)}$, which will generally detect broader features as edges and corners), it is important to note that this SSAE approach is fundamentally different from existent hand-crafted methods that rely on low-level image information such as color, edge cues or texture, as is the case in the texton-based approach presented in 4.1.1.2, since the filters being applied here are learned by the network instead of corresponding to a pre-defined fixed set, as was the case for the filter bank used for the previous approach [14].

4.1.2 Classification

The task of classification is to find a rule, which, based on external observations, assigns an object to one of several classes [70]. In the simplest version, as considered in this work, there are only two classes, hence corresponding to a binary classification problem.

In the supervised machine learning methods proposed, two classifiers were employed, namely support vector machines and naive Bayes. Regarding the SVM algorithm, presented in section 4.1.2.1, its choice was due to the fact that, as previously mentioned, it is known to perform well with high dimensional data, which is the case with the FDG-PET datasets considered for this work. Concerning naive Bayes, the fact that its computational cost can be considerably lower and that its intrinsic assumption leads to having the order in which the features are fed into the classifier be irrelevant, also raises significant interest for the problem at hand, as the images are not registered and so the features associated to the discriminative areas may indeed be encountered in a different order. Regarding the deep learning strategy also considered, a stacked sparse autoencoder, a softmax classifier was applied at the output layer, as described in section 4.1.2.3.

Some considerations regarding training and testing the classifier and evaluating the model performance, taken into account for the experimental setup of this thesis, are then presented in section 4.1.2.4.

4.1.2.1 Support vector machines

Support vector machines (SVMs) are a set of methods for supervised learning, applicable to both classification and regression problems [70]. Regarding the classification problem, the algorithm is based on the assumption that the training set, defined as $T = \{(x_i, y_i), i = 1, ..., n\}$ (where each $x_i \in \mathbb{R}^p$

corresponds to a training pattern, and $y_i \in \{-1, 1\}$ to the class label, in this case referring to a binary classification task) is separable by a hyperplane [70],[71]. Although the simplest version of the method consists of applying it to a set of linearly separable data, further extensions enable it to be applied to a non-linearly separable dataset and also to consider non-linear SVMs, where each of this cases will be detailed in sections 4.1.2.1.1, 4.1.2.1.2 and 4.1.2.1.3, respectively.

To recall the definition of a hyperplane in $\mathbb{R}^p$, this is given by [71]:

$$x \cdot w + b = 0 \tag{4.31}$$

where $w \in \mathbb{R}^p$ corresponds to the normal vector, orthogonal to the hyperplane, and $b \in \mathbb{R}$ to the offset, which, in the case where the Euclidean norm $\|w\|_2 = 1$, corresponds to the signed distance to the origin.

The general idea of the method is then to maximize the distance between the decision boundary (the referred hyperplane) and the closest training patterns (designated as the support vectors). For class $+1$ (or -1) this distance is named as d_+ (or d_-), so that the margin which the method aims to maximize is given by $d = d_+ + d_-$ [71]. It should be noted that this maximum-margin hyperplane may lie in a transformed input space (the feature space), in case non-linear SVMs are applied [70]. Since only the support vectors are required to define the separating hyperplane (which in this case separates the negative from the positive examples), but at the same time these can only be identified after knowing how to formulate the decision boundary (being defined as the closest data points to it), obtaining the solution hyperplane isn't completely straightforward. Instead, the associated parameters are derived from a quadratic programming optimization problem [70].

4.1.2.1.1 Linearly separable data

In the simplest case, where the training patterns can indeed be separated by a hyperplane, the formulation of the SVM algorithm is built on the assumption that the training data satisfies the following constraints [71] :

$$x_i \cdot w + b \geq +1 \quad \text{for} \quad y_i = +1 \tag{4.32}$$

$$x_i \cdot w + b \leq -1 \quad \text{for} \quad y_i = -1 \tag{4.33}$$

using the same nomenclature as in 4.1.2.1, and which is equivalent to:

$$y_i(x_i \cdot w + b) - 1 \geq 0 \ , \quad \forall i \tag{4.34}$$

Considering the support vectors for each class, where the equalities 4.32 and 4.33 (respectively) hold, one can conclude that the distance from these hyperplanes to the origin is given by (respectively) $\frac{|1-b|}{\|w\|}$ and $\frac{|-1-b|}{\|w\|}$, so that both $d_+ = d_- = \frac{1}{\|w\|}$ which results in a margin $d = \frac{2}{\|w\|}$, which the algorithm will aim at maximizing, again subject to the constraints given by 4.34 [71]. This is also equivalent to minimizing

$\|w\|$, or alternatively $\frac{1}{2}\|w\|^2$, so that the optimization problem to be solved can be formulated as [70]:

$$\min \frac{1}{2}\|w\|^2 \ , \quad s.t. \ y_i(x_i \cdot w + b) - 1 \geq 0 \ , \ \forall i \tag{4.35}$$

In order to solve the quadratic optimization problem with constraints given by 4.35, a simple way comes from switching to a Lagrangian formulation of the problem. In fact, not only this allows for the constraints to be placed on the Lagrange multipliers themselves, which will be much easier to handle, but also for having the training data only appear in the form of dot products between vectors, which is a crucial property of SVMs that will allow for its generalization to the case presented in section 4.1.2.1.3. Indeed, this results in:

$$w = \sum_{s \in S} \alpha_s y_s x_s \tag{4.36}$$

$$b = y_s - \sum_{m \in S} \alpha_m y_m (x_m \cdot x_s) \tag{4.37}$$

where these Lagrange multipliers, α_i, satisfy the Karush-Kuhn-Tucker (KKT) condition [72] and using the notation where $s \in S$, the subset of all training patterns that corresponds to the support vectors. Alternatively, the mean value of all resulting b values computed can be taken as the offset b, in this case being given by:

$$b = \frac{1}{\#S} \sum_{s \in S} \left[y_s - \sum_{m \in S} \alpha_m y_m (x_m \cdot x_s) \right] \tag{4.38}$$

where $\#S$ corresponds to the cardinality of the set of support vectors [70],[71]. Classification of the data is then performed as:

$$\hat{y} = f(x) = sign(x \cdot w + b), \ \ \hat{y} \in \{-1, 1\} \tag{4.39}$$

where $\hat{y}$ corresponds to the predicted class label [70],[71]. Moreover, besides the decision, the SVM algorithm also provides a score in the interval $]-1, 1[$ [72].

4.1.2.1.2 Non-linearly separable data

As mentioned, the SVM algorithm can also be extended to deal with non separable data, as is usually the case in more complex classification problems. The general idea is, thus, to allow data points on the wrong side of the hyperplane, provided that they suffer a penalty [70],[71],[73]. Equivalently, the goal is to relax the hard margin constraints 4.32 and 4.33, although introducing a further cost for doing so. The referred equations can then be reformulated by introducing positive slack variables $\xi_i, i = 1, ..., n$, which are defined in such a way that $\xi_i = 0$ if no margin violation occurs and $\xi_i > 0$ if he i-th data point is misclassified [70],[71],[73]. Equations 4.32 and 4.33 then become:

$$x_i \cdot w + b \geq +1 - \xi_i \quad \text{for} \ \ y_i = +1 \tag{4.40}$$

$$x_i \cdot w + b \leq -1 + \xi_i \quad \text{for} \ \ y_i = -1 \tag{4.41}$$

or, equivalently:

$$y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0 \ , \ \ \forall i \tag{4.42}$$

having $\xi_i \geq 0 \ , \ \forall i$. This case is depicted in Figure 4.7, retrieved from [71].

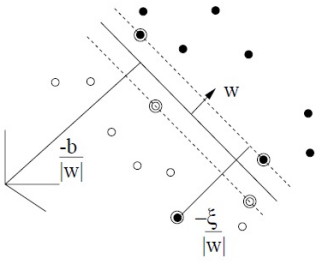


Figure 4.7: Representation of linear separating hyperplanes for the non-separable case. The decision boundary is represented by the solid line and the support vectors are circled. As illustrated, in this example one data point was allowed on the wrong side of the hyperplane.

The optimization problem must then also take the soft margin penalty term into account, as given by $C \sum_{i=1}^{n} \xi_i$, so that 4.35 is reformulated to [70],[71]:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \ , \quad s.t. \ \ y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0 \ , \ \ \forall i \tag{4.43}$$

Once again switching to the Lagrangian formulation, one obtains:

$$w = \sum_{s \in S} \alpha_s y_s x_s \tag{4.44}$$

while parameter b can be obtained in an analogous way to what was presented in the previous section (considering, or not, the mean value from all b parameters computed in this way) [71]. The same rule for data classification as presented for the linearly-separable case is also employed here, as given by:

$$\hat{y} = f(x) = sign(x \cdot w + b), \ \ \hat{y} \in \{-1, 1\} \tag{4.45}$$

Naturally, there is a trade-off between soft and hard margins, as by choosing the former some errors in the training set are allowed, which wouldn't occur in the latter case; however, this can lead to a better generalization of the model and hence a better performance when evaluated on an independent test set, preventing overfitting, so that parameter C (being a hyperparameter) should thus be tuned accordingly - section 4.1.2.4 will go into more detail regarding model generalization and performance [73].

4.1.2.1.3 Non-linear SVMs

A further extension of the SVM algorithm consists of generalizing it to the case where the decision boundary can't be synthesized as a linear function of the data. One way to achieve this consists of mapping the training data, $x \in \mathbb{R}^p$, from the input space to a higher dimension feature space, $\tilde{x} = \Phi(x)$, where the applied transformation must be one that leads the data in the new feature space to be linearly

separable by a hyperplane [70],[71],[73]. This nonlinear mapping is, thus, of the form:

$$\Phi : \mathbb{R}^p \mapsto \mathcal{H} \tag{4.46}$$

where $\mathcal{H}$ corresponds to the new (possibly infinite dimensional) Euclidean space. An example of this case is shown in Figure 4.8, retrieved from [70], where the associated mapping to a 3-dimensional space led, indeed, the transformed data to become linearly separable.

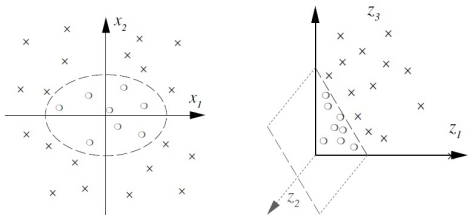


Figure 4.8: Example of a nonlinear decision boundary in the 2D input space and corresponding linear separating hyperplane in the transformed 3D feature space.

As highlighted in section 4.1.2.1.1, however, solving the quadratic optimization problem to find the separating hyperplane doesn't require that the training data is known in itself; on the contrary, only the inner product between the input patterns is needed for that computation [70],[71],[72],[73]. As a consequence, considering this nonlinear mapping to a higher dimension feature space, the algorithm would once again only depend on the data through dot products, and so if it was possible to obtain the result of that inner product through a kernel function, as given by:

$$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \tag{4.47}$$

explicit knowledge of the transformed vectors wouldn't be required and all computations performed to solve for the decision boundary would again only depend on that inner product [70],[71],[73]. This strategy is commonly designated as the kernel trick, and, although the feature space in which SVM is applied may even be infinite dimensional, since the computation of the high dimensional features in itself isn't required, the computational cost is similar to that associated to training the unmapped data. Moreover, the considerations in the previous sections still hold, since this will still be a linear separation (although in the feature space), apart from slight differences in terms of the optimization problem's formulation (derived from replacing the training patterns in the input space, given by x_i and x_j, by the transformed ones, $\Phi(x_i)$ and $\Phi(x_j)$) and in terms of the rule for data classification, which can be formulated as:

$$\hat{y} = f(x) = sign\left(\sum_{s \in S} \alpha_s y_s K(x_s, x) + b\right), \;\; \hat{y} \in \{-1, 1\} \tag{4.48}$$

where x_s once again correspond to the support vectors, and the independent vector x to the pattern to be classified [70],[71],[72],[73].

Several examples of allowed kernel functions (meaning, that meet the condition 4.47) have been proposed, including linear, radial basis function (RBF) and generalized histogram intersection (GHI),

which have been formulated according to equations 4.49, 4.50 and 4.51, respectively:

$$K_{Linear}(x_i, x_j) = x_i^T x_j \tag{4.49}$$

$$K_{RBF}(x_i, x_j) = e^{-\frac{1}{2\sigma^2}\|x_i - x_j\|^2} \tag{4.50}$$

$$K_{GHI}(x_i, x_j) = \sum_{k=1}^{p} min(|x_{ik}|^\beta, |x_{jk}|^\beta) \tag{4.51}$$

where, in what refers to 4.51, x_{ik} corresponds to the k-th component of the i-th training sample [12],[70], [71], [74]. While the linear kernel corresponds to the simplest version of the SVM algorithm as presented in 4.1.2.1.1, RBF and GHI are more complex, also including hyperparameters in its formulation, namely σ and β, respectively, which must be specified or else tuned namely through nested cross-validation as will be described in section 4.1.2.4 [12],[70],[74].

4.1.2.2 Naive Bayes

Another classification algorithm that has proved to be very effective in a number of applications, including medical diagnosis, having indeed exhibited a competitive predictive performance with regard to the state of the art, is the naive Bayes classifier [75],[76]. As the name indicates, the formulation for this algorithm derives from the simple theorem of probability known as Bayes law, as given by [77]:

$$P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)} \tag{4.52}$$

where $x \in \mathbb{R}^p$ is a feature vector and ω_i is a particular class, having $\Omega = \{\omega_0, .., \omega_{K-1}\}$ correspond to the set of K classes considered for the classification task. Using Bayes law, one can thus compute the conditional probability that the input pattern belongs to class ω_i, knowing that it has feature vector x (or, equivalently, the *a posteriori* distribution of the classes after observing the feature vector x), given by $P(\omega_i|x)$, from the distribution of the feature vector x associated to class ω_i, given by $P(x|\omega_i)$, and the *a priori* distribution of the classes (prior to taking the observations into account), given by $P(\omega_i)$. $P(x)$ corresponds, in turn, to a normalization term, as given by [77]:

$$P(x) = \sum_{\omega \in \Omega} P(x|\omega)P(\omega) \tag{4.53}$$

Assuming that $P(\omega_i|x)$ can be computed exactly through this equation (if the distributions $P(x|\omega_i)$ and $P(\omega_i)$ are known), the classification can be done in an optimal way, in the sense that the probability of decision error can be minimized, by assigning the input pattern with feature vector x to the class ω_i for which this *a posteriori* probability $P(\omega_i|x)$ is highest [75],[77]. This corresponds, precisely, to the framework used in the Bayes classifier, where the predicted class label, $\hat{y}$, is given by [75],[77]:

$$\hat{y} = arg \max_{\omega \in \Omega} P(\omega|x) \tag{4.54}$$

Nonetheless, this is usually not the case, so that $P(\omega_i|x)$ must be estimated from the data. Particularly,

using a training set $T = \{(x_i, y_i), i = 1, ..., n\}$, composed of a set of n training patterns and respective class labels, with $x_i \in \mathbb{R}^p, \forall i$ and $y_i \in \Omega = \{\omega_0, .., \omega_{K-1}\}$, $P(x|\omega_i)$ and $P(\omega_i)$ can be estimated and combined to get an estimate of the desired *a posteriori* probability. However, this can be a very difficult problem in case the feature vector x contains many features [77]. This computation is then rendered feasible by making a strong independence assumption, giving rise to the naive Bayes classifier. Particularly, it assumes that all the features are conditionally independent given the value of the class, so that we only need to estimate the conditional distribution of each individual feature [75],[76],[77]. This can then be formulated as:

$$P(x_1, ..., x_p|\omega_k) = \prod_{i=1}^{p} P(x_i|\omega_k) \tag{4.55}$$

which can be replaced in equation 4.52, holding:

$$P(\omega_k|x) = \frac{\prod_{i=1}^{p} P(x_i|\omega_k)}{P(x)} \times P(\omega_k) \tag{4.56}$$

Using this formulation, feature vector x can then be classified as within the class ω_k for which this probability is higher, as mentioned. It should also be noted that the computation of the normalization term $P(x)$ is not required for classification, as it isn't class dependent. Even though the conditional independence assumption imposed in the naive Bayes classifier is unrealistic, turning into a suboptimal classifier in case it is not true, it often leads to good results, hence turning it into a an effective classification algorithm [75]. Furthermore, the fact that the features are assumed to be conditionally independent also ensures that the algorithm's performance is completely independent of the order in which the input features are fed into it. In line with these considerations, the naive Bayes classifier was, thus, also tested in the experimental part of this thesis.

4.1.2.3 Softmax Classifier

Concerning the softmax classifier, it is a supervised model typically used in the final layer of deep neural network architectures, namely stacked sparse autoencoders, as described in 4.1.1.3. It generalizes the logistic regression function to multi-class classification problems, so that its application at the output layer, l, results in:

$$h_i^{(l)} = \frac{e^{W_i^{(l-1)} h^{(l-1)} + b_i^{(l-1)}}}{\sum_j e^{W_j^{(l-1)} h^{(l-1)} + b_j^{(l-1)}}} \tag{4.57}$$

where $W_i^{(l-1)}$ and $b_i^{(l-1)}$ correspond to the weight and bias terms associated to the connection between the previous and unit i in the current layer, $h^{(l-1)}$ corresponds to the activation map at the output of the previous layer and $h_i^{(l)}$ corresponds to the activation of the i-th unit of the output layer [47]. Each unit in layer l is associated to a class, so that the total number of units in this final layer is the same as the number of classes considered for the classification problem [47]. As the softmax function corresponds to a normalized exponential function, so that the sum of its output values over all classes is 1, the output value $h_i^{(l)}$ can thus be used as an estimator of the conditional probability that the input pattern belongs to class ω_i, $P(y = \omega_i\,|x)$, where y is the associated label of input data vector x, using the same notation as

in the previous sections, again having $\Omega = \{\omega_0, .., \omega_{K-1}\}$ correspond to the set of K classes considered for the classification task [47]. The softmax classifier is, thus, able to attribute the input feature vector to the class ω_i for which $h_i^{(l)}$ is the highest [47].

In the case of a binary classification problem, as is the case in the current thesis, the expression of the softmax function is again simplified to hold the same formulation as in 4.14, for the logistic sigmoid function [14]. Again considering its application at the output layer, added on top of a SSAE architecture, and particularly considering one with two hidden layers, as presented in 4.1.1.3.3, the resulting secondary features can be treated as raw input to that classifier, training it to map secondary features to the respective data labels [67], having the softmax classifier be formulated as:

$$f_{W^{(3)}}(z) = \frac{1}{1 + exp(-W^{(3)^T} z)} \tag{4.58}$$

where $f_{W^{(3)}}$ is the sigmoid function with parameters $W^{(3)}$, corresponding to the weights entering the output, softmax layer, (in this case, being the fourth layer) and z corresponds to the learned, high-level structure information, namely the secondary feature representation, $h^{(3)}$ [14]. The softmax classifier hence produces a value between 0 and 1 that can be interpreted as the probability of the input feature vector (in the current work, corresponding to a constructed 3D image patch) belonging to the class labeled as 1 or 0 [14].

Once again considering the SSAE strategy applied in this work, the softmax classifier is trained using the backpropagation algorithm, namely using the gradient descent approach (amongst other that have proved to be more effective [78]), so that the optimal weights $W^{(3)}$ can be obtained [14]. The loss function typically used in this case corresponds to the cross-entropy loss, which can be formulated as:

$$J(W, b) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} \Big[y_i \; log(h_{W,b}(x^{(i)})_j) \Big] \tag{4.59}$$

where y_i corresponds to the true class label for input pattern $x^{(i)}$, n corresponds to the total number of training samples, K corresponds to the total number of classes considered, and $h_{W,b}(x^{(i)})_j)$ corresponds to the activation at the output (softmax) layer for unit j, for the input pattern $x^{(i)}$ [13]. Once again, in binary classification problems, this expression can be simplified, and formulated as:

$$J(W, b) = -\frac{1}{n} \sum_{i=1}^{n} \Big[y_i \; log(h_{W,b}(x^{(i)})) + (1 - y_i) \; log(1 - h_{W,b}(x^{(i)})) \Big] \tag{4.60}$$

where $h_{W,b}(x^{(i)})$ corresponds to the activation at the output (softmax) layer for the input pattern $x^{(i)}$.

4.1.2.4 Model Selection and Performance Evaluation

The first stage of a supervised learning problem, particularly a classification task, consists of training the model. This requires a training set, as represented by $T = \{(x_i, y_i), i = 1, ..., n\}$, where each $x_i \in \mathbb{R}^p$ corresponds to a training pattern (having p correspond to the number of input features considered for each sample), y_i to the class label (in this case having $y_i \in \{0, 1\}$, as it refers to a binary classification

task) and n to the cardinality of the training set. Being trained with these input patterns, the output of the learned system almost always corresponds to the desired output if tested on the same training set, causing the evaluation of model performance to be too optimistic, and hence unreliable.

In order to measure the ability of the system to predict the outcome $\hat{y}, \hat{y} \in \{0, 1\}$, in a new case (meaning the generalization of the system), it is thus mandatory that this evaluation is made on a new, independent, test set [79]. However, when the model includes hyperparameters to be chosen, these must be tuned with resource to yet a new independent data set, named as the validation set. Thus, while the training set is used to learn the model, its hyperparameters should be selected according to the empirical risk obtained in the validation set, as given by:

$$\mathcal{R} = \frac{1}{n'} \sum_{i=1}^{n'} L(y'_i, f(x'_i)) \tag{4.61}$$

where $T' = \{(x'_i, y'_i), i = 1, ..., n'\}$ corresponds to that independent validation set of size n', where the feature vectors x'_i and class labels y'_i are of the same form as the ones in the training set, having $f(x'_i) = \hat{y_i}'$ correspond to the predicted outcome for x'_i. Function L corresponds, in turn, to the loss function, which, being formulated as a binary loss, so as to attribute an equal penalization to false positives and false negatives, is given by:

$$L(y'_i, \hat{y_i}') = \begin{cases} 0, & y'_i = \hat{y_i}' \\ 1, & y'_i \neq \hat{y_i}' \end{cases} \tag{4.62}$$

Once again, this is done since an estimate of the model performance with a given choice of hyperparameter on the training set would be too optimistic, with the selected model being too adjusted to the training set (corresponding to an overfitting phenomenom) possibly even modelling noisy data, and exhibiting a poor generalization ability when evaluated in the independent test set.

So, by separating all the data in these three disjoint sets (training, validation and test), it is possible to learn the model in the training set using each choice of the hyperparameter value at a time and select the best one from the resulting performance in the validation set, and finally, using this final learned classifier its generalization ability can be measured in the test set. There is, however, a drawback to this procedure, since it requires a lot of data in order to be able to discard part of it at the training stage [79]. One alternative thus consists of nested cross-validation (illustrated in figure 4.9), built from the simpler version of cross-validation, but in this case including the choice of hyperparameters as required namely when using the SVM algorithm and its kernel functions, as used for the purpose of this work, according to the experimental setup to be presented in 4.2.2 [79].

Figure 4.9: Illustration of the data partitioning procedure for nested cross-validation.

Using nested cross-validation, the initial dataset is first partitioned into k disjoint sets, for instance in

$k = 10$ folds. As the name indicates, this method includes an outer and an inner loop where at each iteration of the former one of the k folds is used as the test set, which thus rotates k times, and, for each of those, the inner loop is cycled through for model selection. In detail, at each iteration of this nested loop, a hyperparameter value is chosen and yet another nested loop is cycled through. There, from the remaining $k - 1$ folds, $k - 2$ are used for training and 1 as the validation set, and the latter rotates $k - 1$ times. For each iteration of this innermost loop, a performance score is obtained for the validation set being used and, once all folds are considered for this purpose, the average performance score for that chosen hyperparameter is obtained. After computing the corresponding scores for all possible hyperparameter values, the best configuration is selected and the model is then learned with the chosen parameters, using all feature vectors other than the ones in the test set for that iteration of the outer loop. The classifier is then applied on the test set to evaluate the system's performance, which, as mentioned, rotates k times, so that the final score is then a combination of the k models, particularly given by its average performance scores.

Several metrics can be considered to evaluate model performance, and particularly for classification tasks, these include accuracy, sensitivity, specificity, as given by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.63}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.64}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.65}$$

where TP stands for the number of true positives, TN for the number of true negatives, FP for the number of false positives and FN for the number of false negatives.

The three different metrics were indeed used in this work to allow for establishing a comparison between the different methods explored where, adapting these formulations to the binary classification problems here considered, a true negative corresponds to a correctly diagnosed cognitively normal subject, and a false negative to a subject that is misclassified as cognitively normal, despite belonging to the remaining class (either AD or MCI, depending on the binary classification task), which would correspond to the ground truth. An analogous reasoning can be applied for the true positive and false positive cases, corresponding to cases where the subject is correctly diagnosed as having AD (or MCI, accordingly) or misclassified as such despite being a healthy control subject, respectively. Hence, the accuracy informs of the number of patients that were correctly diagnosed, regardless of the class, while the sensitivity informs about the number of AD (or MCI) bearing subjects that were correctly classified as such (so that a high sensitivity implies a small amount of AD or MCI cases that were disregarded due to being labeled as cognitively normal) and the specificity informs about the number of cognitively normal subjects (so that a high specificity implies a small amount of cases being false alarms, due to an incorrect positive diagnosis). Since all these metrics provide relevant and complementary information, all three were thus considered for performance evaluation.

4.2 Implementation

4.2.1 Datasets

To evaluate the accuracy and robustness of the presented methods, three distinct datasets were considered for each class, two of which consisted of real FDG-PET brain images having, or not, already undergone image registration, and the remaining one having been artificially generated from those.

Regarding the two real datasets, these were obtained from the ADNI database, which stands for Alzheimer's Disease Neuroimaging Initiative. Initially launched in 2004, this project consists of a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease [80]. Amongst other neuroimaging modalities such as MRI, FDG-PET data collections are available in this database, including both the original acquired brain images and the pre- and post-processed ones, to minimize differences between images and allow for voxel-wise comparisons. Regarding the second, four types of processed PET images exist in ADNI, the latter of which corresponds to the dataset used in this work as the one prior to image registration, consisting of co-registered, averaged, standardized images and voxel sizes (reoriented into a 160x160x96 voxel image grid) as well as with uniform resolution [80]. More specifically, after dynamically co-registering separate frames from the raw images to one another to lessen the effects of patient motion and averaging the resulting image dataset, the resulting baseline PET scan for each patient had already been reoriented (for the anteroposterior axis of the subject, as represented in Figure 4.10, retrieved from [81], to be parallel to the AC-PC line), intensity normalized and smoothed to a uniform standardized resolution [80]. Nonetheless, it's important to mention that no non-linear warping or even linear scaling had been applied to these brain images [80]. As for the corresponding dataset after registration, data from the same subjects was retrieved, in this case in the form of post-processed images. Indeed, after the same pre-processing procedure as supra-mentioned, the selected images had also been non-linearly warped to the Talairach brain atlas, resulting in a new voxel image grid of 128x128x60 [80].

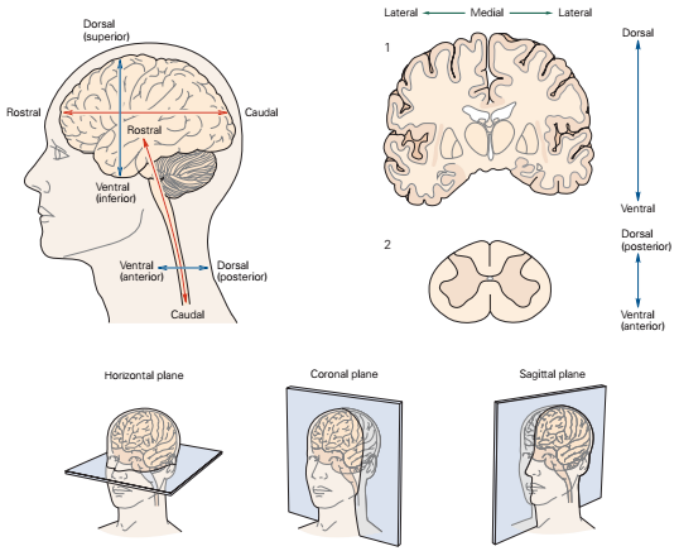


Figure 4.10: Representation of the three major anatomical axes and planes.

In relation to the artificially generated dataset, it derived from applying different 3D affine transforms to each of the registered images, particularly 3° incremental rotations along the z-axis (inferior/superior, in agreement with the anatomical axes illustrated in Figure 4.10) ranging from 3° to 309°. To clarify, the x-axis and y-axis correspond, in turn, to the anteroposterior and mediolateral axes, as also represented in Figure 4.10. The choice of including this dataset in addition to the two real ones was made to allow for evaluating the robustness of the implemented methods in images that were significantly and manifestly disaligned, in contrast to the much slighter deviations observed within the pre-processed dataset.

As mentioned, all three datasets refer to the same subjects, who in turn belonged to one of three classes, either being cognitively normal, diagnosed with MCI or diagnosed with AD. A sample image for each of the three classes considered and for each dataset is presented in Figure 4.11. Subject selection was performed by imposing class dependent restrictions based on their, also available, neurological assessment and in particular regarding the Clinical Demetia Rating (CDR). A score of 0, 0.5 and 0.5 or higher was thus imposed for healthy controls, MCI patients and AD patients, respectively, according to what was presented in section 2.3.2 regarding this test. This resulted in a dataset composed of 53, 80 and 50 subjects, respectively, regarding whom important clinical and demographic information is summarized in Table 4.1.

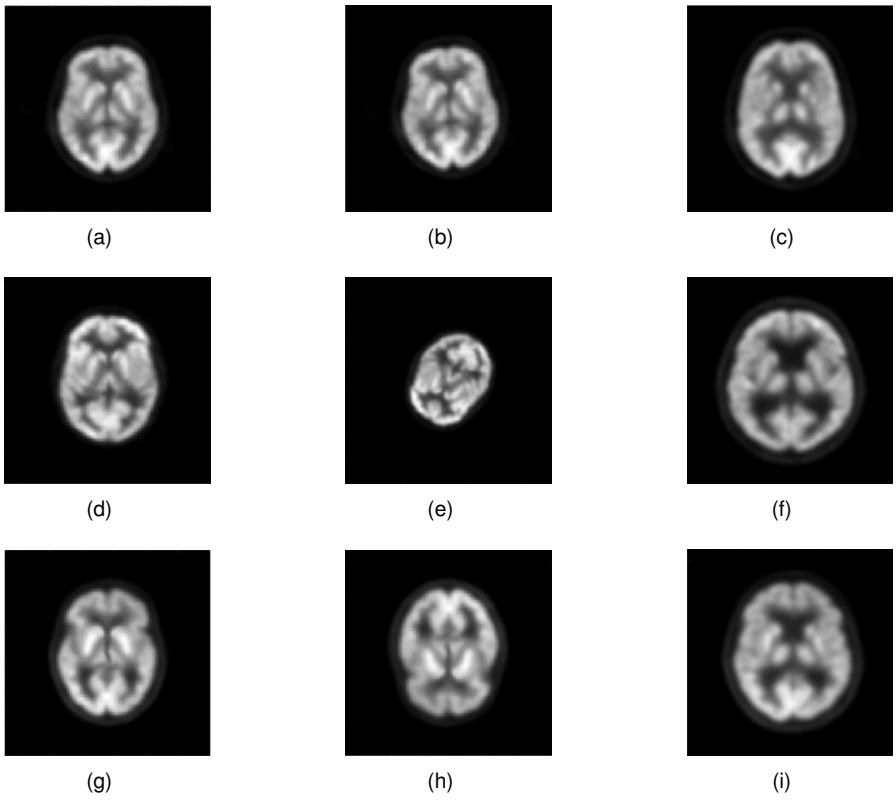


Figure 4.11: Sample images for each class and dataset. Figures (a) to (c) refer to a subject in the CN class, figures (d) to (f) to a subject in the MCI class and figures (g) to (i) to a subject in the AD class, respectively for the registered, artificially generated and non-registered datasets.

Table 4.1: Characteristics of each group of subjects. Format: Mean (Standard Deviation).

Attributes	AD	MCI	CN
Nº of subjects	53	80	50
Age	75.9 (6.7)	74.5 (6.6)	75.4 (4.9)
Sex (% of Males)	56.6	72.5	56.0
MMSE	23.2 (2.1)	27.3 (1.7)	29.1 (0.9)

4.2.2 Experimental setup

For the texton-based approach, the filter bank that was used was a further extension to the 3D version of MR8, as introduced in section 4.1.1.2.1, composed by 5 types of filters at several scales and orientations, particularly Gaussian, Laplacian of Gaussian (LoG), edge, bar and plane filters. All 3D filters were included at 3 triplets of scales: $(\sigma_x, \sigma_y, \sigma_z) = \{(1,1,1);(2,2,2);(4,4,4)\}$ for the rotationally invariant Gaussian and LoG filters (hence the difference from the 3D approach presented in section 4.1.1.2.1, derived from [12], where a single scale was used for these two isotropic filters, namely $\sigma = 2$); $(\sigma_x, \sigma_y, \sigma_z) = \{(1.5,1.5,0.5);(3,3,1);(6,6,2)\}$ for the edge and plane filters; and finally $(\sigma_x, \sigma_y, \sigma_z) = \{(0.5,0.5,1.5);(1,1,3);(2,2,6)\}$ for the bar filter. According to section 4.1.1.2.1, for 3D extension, the last 3 types of filters were replicated with multiple orientations through systematic sampling of angles θ and φ, defined around the z and y axis, respectively. Indeed, θ was sampled in a range from $\frac{\pi}{6}$ to $\frac{5\pi}{6}$, while φ was sampled in a range from 0 to $\frac{11\pi}{6}$, both of which at a $\frac{\pi}{6}$ step. Also considering the initial orientation, this resulted in a total of $(3 \times 5 \times 12 + 3) \times 3 + 3 + 3 = 555$ filters and subsequent filter responses. However, only 9 filter responses were used, since as to achieve rotation invariance only the maximum across all orientations was kept for each scale triplet regarding the edge, plane and bar filters. Adding to the 6 filter responses that derived from the Gaussian and LoG filters resulted in 15 filter responses and thus in a 15-dimensional filter response vector for each voxel of each brain image.

Regarding the classifiers, and considering the small size of the dataset used for each binary classification task (103 images for the CN vs. AD problem and 130 for CN. vs MCI), a nested cross-validation procedure was used to tune the parameters associated to each SVM kernel used, namely linear, RBF or GHI. For each task, the data was appropriately partitioned (avoiding class imbalance) in 10 folds for the outer loop and in 5 folds for the inner one. The size of the test set was, then, either 10 or 11 for CN vs. AD and 13 for CN vs. MCI. From the remaining 92 or 93 subjects for CN vs. AD, the validation set comprised either 18 or 19, while for the remaining 117 subjects for CN vs. MCI either 23 or 24 constituted the validation set. A grid-search procedure was then used to tune the hyperparameters, as suggested in [82]. This was done with resource to the *LIBSVM* toolbox [83], compatible with *MATLAB R2016a*, the software used throughout all the course of this work. The allowed values were such that $C \in \{2^{-10}, 2^{-8}, ..., 2^{10}\}$, corresponding to the soft margin of the SVM algorithm, $\beta \in \{0.1, 0.4, ..., 1.9\}$ for the GHI kernel, and $\gamma \in \{2^{-11}, 2^{-9}, ..., 2^{3}\}$ for the RBF kernel where, using the same notation as in the *LIBSVM* toolbox, γ corresponds to $\frac{1}{2\sigma^2}$ in the expression presented in 4.50. This toolbox was used both for training and testing the SVM classifier. As for the naive Bayes algorithm, to account for the

possible occurrence of very small texton frequencies in the extracted histograms, instead of using the probability of occurrence of each of those features in either class, as formulated in expression 4.56, in section 4.1.2.2, the classification was performed using the logarithm of its value. Moreover, to ensure that neither of those probabilities would equal 0, as this would result in a undetermined $log(0)$ computation, Laplace (additive) smoothing was introduced. The logarithm of $P(x_i|\omega_k)$, having x_i correspond to feature i in the feature vector to be classified, was then replaced by $log(\frac{\#x_{i,k}+1}{\#x_{\omega_k}+N_{tot}})$, where $\#x_{i,k}$ corresponds to the number of occurrences of feature x_i in class ω_k, $\#x_{\omega_k}$ corresponds of the total number of features (or equivalently, textons) in class ω_k, both referring to the training set, and N_{tot} corresponds to the total number of textons in the dictionary.

Concerning the deep learning strategy used in this work, a stacked SAE, it was trained with the *Neural Network Toolbox* provided by *MATLAB R2016a*. For this purpose, two autoencoders (of hidden sizes 100 and 50, respectively) were used, followed by a final softmax layer. The initial autoencoder was trained during a maximum 400 epochs and subjected to two regularization techniques, namely an L2 regularizer for the weights of the network (typically with a very small impact, in this case having the associated coefficient been set at 0.004) and a sparsity regularizer, to enforce the sparsity constraint on the output from the hidden layer (having the impact parameter been set at 4). The sparsity parameter ρ, as defined in 4.1.1.3.2, was set at 0.15 (where a lower value indicates that each neuron in the hidden layer only gives a high output for a small number of training examples, as mentioned). Sequentially, the second autoencoder was trained for a maximum of 100 epochs and using analogous parameters, respectively with the values of 0.002, 4 and 0.1. The output of the hidden layer of the second autoencoder was then used to train the softmax layer during a maximum number of epochs of 400, so as to classify the resulting 50-dimensional feature vectors. To improve the performance of the constructed stacked neural network, it was then fine tuned, by performing backpropagation on the whole multilayer network, during a maximum number of epochs of 100. Furthermore, instead of using the gradient descent for backpropagation as presented in 4.1.1.3, a scaled conjugate gradient algorithm (SCG) was used, as proposed in [78], as it has proved to be a faster and more effective method. Indeed, unlike the gradient descent which uses a linear approximation of the error function (which is the main reason why the algorithm often shows poor convergence) and a constant step size (which in many cases is inefficient and makes the algorithm less robust), the SCG algorithm uses second order information as well as adaptive step sizes, in a fully automated way, including no user-dependent parameters. While including a momentum term in the generic gradient descent algorithm is an attempt to force the algorithm to use second order information from the network, it still doesn't provide a considerable speed up in the algorithm, and causes the algorithm to be even less robust, because of the inclusion of another user dependent parameter, the momentum constant [78]. Moreover, contrary to other algorithms that are usually adopted for this purpose such as BFGS (which stands for the Broyden-Fletcher-Goldfarb-Shanno algorithm), SCG provides the advantage of avoiding the time consuming line-search, done in each iteration of BFGS to determine an appropriate step size [78]. Considering these advantages, the SCG algorithm was applied for training the SSAE neural network. The performance function used for the two stacked autoencoders consisted of the mean squared error function adjusted for training a sparse autoencoder, according to

equation 4.29, while the cross-entropy function, as defined in 4.60, was used for supervised learning, when training the final softmax layer as well as for the fine tuning procedure.

Further details regarding the implementation and performance evaluation of each of the applied methods following each feature extraction approach considered, particularly using the whole brain or patches of the brain, will be presented in the following section.

4.2.3 Proposed approaches

The explored methods were evaluated on two binary classification tasks, namely CN vs. AD and CN vs. MCI, and, for each of these, as referred in 4.2.1, three datasets were used. Even though the respective images include the area surrounding the brain (as shown in Figure 4.11), since only the regions within it are relevant for classification, the surrounding area can be fully disregarded, allowing not only to exclude non-discriminative voxels, but also to largely reduce the size of the feature vectors. For this purpose, a brain mask was used, so that only the voxels inside this mask were kept and considered for feature extraction. Concerning the registered dataset, this consisted of a pre-defined mask in the Talairach space (represented in Figure 4.12, of the same dimensions as the respective images, namely $128 \times 128 \times 60$), while in the non-registered dataset it was created from the brain scans of all the different subjects. Referring to the latter, this was performed by averaging all brain scans, followed by keeping the voxels for which this average intensity value was above a threshold of 0.5 of the maximum intensity value. A morphological operation was then performed so as to close the mask, filling in small regions that had been excluded by this procedure but that were visually identified as a part of the whole brain area. Regarding the artificially generated dataset, each subject was associated to one brain mask, each generated using the same affine transform that had been used to obtain the brain image for that particular subject, in this case applied on the Talairach brain mask considered for the registered dataset.

(a) (b) (c)

Figure 4.12: Representation of the brain mask applied in the registered dataset. Figures (a), (b) and (c) represent sagittal, coronal and axial sections of the brain, respectively.

Furthermore, for each of these classification problems, two feature extraction approaches were considered, namely using the whole brain, and using patches of the brain. This choice was made considering that, since the purpose of this work was to develop a method that could be applied on a dataset on which image registration had not been performed, neither voxel-based approaches nor regions of interest could be used directly, as the location of the discriminative voxels and regions of interest would slightly vary across all subjects. On the contrary, no prior identification of the discriminative areas had to be performed if the features were extracted from the whole brain, or if the features used for classification

could be obtained from different locations in different images, as enabled by using features extracted from patches of the brain, following appropriate patch selection for each subject considered.

4.2.3.1 Feature extraction from the whole brain

Regarding the classification algorithms applied on features extracted from the whole brain, these were tested in all three datasets, following the subject selection procedure presented in 4.2.1. Both SVMs (with linear, GHI and RBF kernels) and naive Bayes were applied, either using the voxel intensities feature vectors (in that case using only the SVM classifier) or the texton-based approach as previously described. According to the texton-based approach as presented in 4.1.1.2, prior to image classification and to building the dictionary of textons, the images were normalized, as well as the obtained filter responses. Regarding the image normalization pre-processing step, this was performed, in each dataset, over the voxels within the considered brain mask, by subtracting the mean and dividing by the standard deviation. Moreover, several experiments were performed regarding the number of textons that should be used to build the dictionary, for which the sequential k-means algorithm was employed (the original k-means algorithm was also tested, although worse performances were achieved). A number of textons ranging from 15 to 10000 was thus considered.

4.2.3.2 Feature extraction from patches of the brain

Concerning the use of selected patches of the brain for classification, three distinct approaches were considered. Regarding the former, this was performed by comparing the texture information from each patch with that of patches located within defined ROIs of the brain. In the second approach, on the other hand, this was done without taking into account any *a priori* knowledge in terms of which regions of the brain should be considered for the diagnosis; instead, following the texton-based method, the different patches were selected solely based on whether discriminative textons were present in its distribution, which where identified as such by using the mutual information ranking. Finally, to evaluate how these two proposed methods could outperform the simplest and theoretically worse case, consisting of a random choice of patches of the brain, the latter approach was also considered.

In what refers to the segmentation of the brain images in patches, these were constructed only within the brain masks considered for each dataset, using a 75% overlap between patches. This choice in overlap was made considering the trade-off between computational cost and coverage of the different regions of the brain, as a higher overlap would result in an increased number of patches and thus in a higher computational cost when performing patch selection. For each of the proposed methods, since the study of the influence of the size of the 3D patches would further largely increase the computational cost, a choice was made to set this parameter at a fixed volume of $7 \times 7 \times 7$.

4.2.3.2.1 Patches within ROIs

Unike the previous approach, using the whole brain, where the image registration step could be disregarded both when training and testing the classifier, this strategy was considered in the case that this

step would still be performed for the training stage. Although this methodology is not ideal, it still allows for avoiding the image registration step when applying the CAD system to diagnose novel subjects, only requiring it to be performed prior to this step, when learning the model, again saving computational costs. Moreover, the fact that image registration still had to be performed for the purpose of training the classifier, resulted from the way in which the different patches were learned to be labeled as discriminative or not regarding patch selection, prior to the final classification stage using the respective extracted features. Indeed, considering that some regions of interest have been identified in the literature for the purpose of CAD of Alzheimer's disease, as mentioned in sections 2.3.3 and 3.4, a patch was labeled as discriminative if it spanned over one or multiple ROIs, particularly considering the ones identified in [12], as shown in Figure 4.13. These include the lateral temporal, mesial temporal, inferior frontal gyrus and orbitofrontal, inferior anterior cingulate, dorsolateral parietal, superior anterior cingulate, posterior cingulate and precuneous [12].

Furthermore, as the image resolution concerning the non-registered dataset differed from that on the registered one (considering the size of $160 \times 160 \times 96$ instead of $128 \times 128 \times 60$ for the previous one), in this case the training stage could not be performed in the registered dataset. Thus, in the case where the training stage was performed on the registered dataset, conclusions could be drawn from the robustness of the methods explored following this feature extraction approach only by assessing its performance in the registered dataset (hence used for training and testing). To evaluate the performance of the classifier on the real non-registered dataset, on the contrary, the ROIs were manually identified from the ones in the Talairach brain atlas, so as to perform the training stage also using this dataset, although once again patch selection had to be performed for the testing stage.

(a) (b) (c)

Figure 4.13: Representation of the regions of interest in the Talairach brain atlas, as labeled in [12]. Figures (a), (b) and (c) represent sagittal, coronal and axial sections of the brain, respectively.

For the training stage, in the registered dataset, a patch was thus labeled as discriminative if over 90% of its volume was within the reunion of these ROIs (for the respective dataset). A classifier was then trained to learn which patches were discriminative using an SSAE combined with softmax classifier. Following this step, the selected patches were then used for the final diagnosis. This was performed using either a new SSAE combined with an output softmax layer or the histogram of textons extracted from the reunion of all selected patches. Regarding the texton-based approach for the final classification of the selected patches, once again the naive Bayes and SVM algorithm (with linear, RBF and GHI kernels) were used, considering a fixed number of textons of 1250 in the dictionary.

4.2.3.2.2 Patches containing discriminative textons

As mentioned, another method that was explored was to select patches based on the frequency of the most discriminative textons on its histograms. These textons were identified as discriminative on the basis of the mutual information feature ranking, as given by:

$$MI = \sum_{x \in X, y \in Y} p(x,y)\ log\left(\frac{p(x,y)}{p(x)\ p(y)}\right) \tag{4.66}$$

where $p(x,y)$ corresponds to the joint probability function of X and Y and $p(x)$ and $p(y)$ correspond to the marginal probability distribution functions of X and Y, respectively, having, in this approach, y correspond to the class label and x to the frequency of each texton in each subject's brain image. These textons corresponded, in turn, to the ones used to build the dictionary in the previous approach (when performing feature extraction from the whole brain), as presented in 4.2.3.1, particularly to the case where 1250 textons were used.

The textons were then sorted so that the most discriminative ones corresponded to those for which the mutual information ranking was higher. As the number of textons that should be considered as discriminative was unknown (corresponding to a hyperparameter for the model) instead of being fixed at a certain value, it was chosen through a nested cross-validation procedure, as explained in 4.1.2.4. This number varied between 1, 5, 10 or 15, and a patch was thus selected if its distribution contained any of these most discriminative textons.

The final diagnosis was obtained by applying, once again, SVM (with linear, GHI and RBF kernels) or naive Bayes for classification of the features extracted from the selected patches.

4.2.3.2.3 Random patch selection

Finally, to evaluate how the proposed methods could outperform the simplest and theoretically worse case, consisting of a random choice of patches of the brain, the latter approach was also evaluated, selecting a number of patches equal to the number of discriminative patches as identified in the ROI-based patch selection strategy. Particularly, this corresponded to 1004, 1250 and 5836 selected patches for, respectively, the registered, artificially generated and non-registered datasets. Once again, SVM (with linear, GHI and RBF kernels) and naive Bayes classifiers were tested in this case.

4.2.3.3 Summary

A flowchart of all the explored methods is presented in Figure 4.14. As mentioned, the aforementioned approaches were applied equally for both binary classification problems, either between Alzheimer's disease and cognitively normal subjects, or between mild cognitive impairment and cognitively normal subjects, hence following the same pipeline for both tasks. For illustration purposes, the only classification task explicitly represented consists of the former.

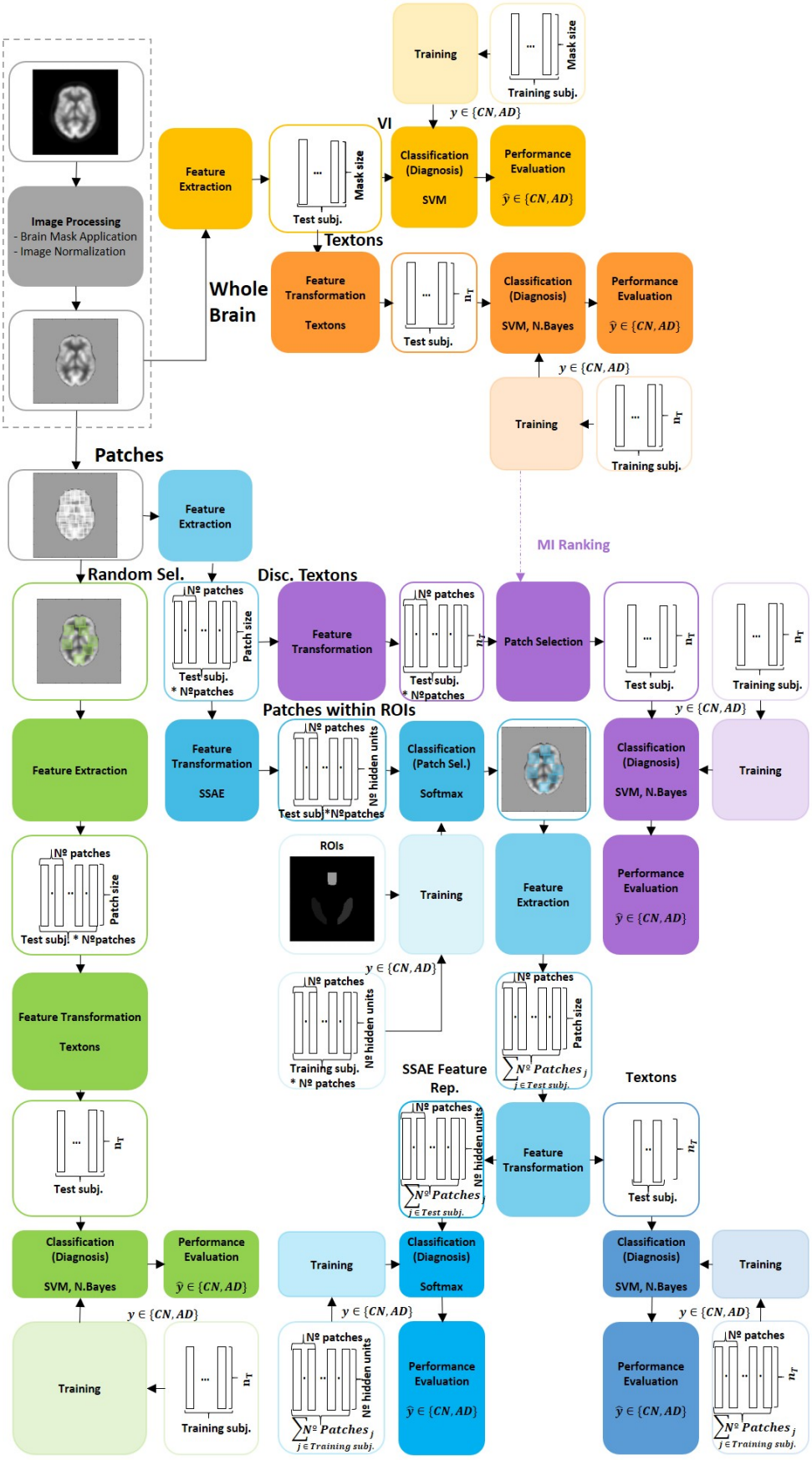


Figure 4.14: Flowchart summarizing the different methods explored.

Chapter 5

Results and Discussion

In this section, the results obtained for each binary classification problem (cognitively normal vs. Alzheimer's disease and cognitively normal vs. mild cognitive impairment) are presented and discussed. As mentioned in 4.2.3, two feature extraction approaches were tested on each classification task (one using the whole brain images, and the second using selected patches) and applied on the different datasets previously described (namely registered, generated and non-registered), so that the following subsections will go into detail regarding the results obtained through each of these methods.

5.1 Cognitively normal vs. Alzheimer's disease

5.1.1 Feature Extraction From The Whole Brain

Regarding the classification of features extracted from the whole brain, all the three datasets introduced in 4.2.1 were used. Concerning the two retrieved from ADNI [80] (registered and non-registered), both the raw voxel intensities and histogram of textons were considered, while, in case of the dataset that was artificially generated, only the second approach was tested, as the images thus produced and respective brain masks had different dimensions, resulting in a different number of features which could not be fed into the classifier without further transformation, which on the other hand was made possible using the texton-based approach.

The parameters associated with the different classifiers were tuned inside a nested cross-validation procedure, according to 4.2, as shown in Figure 5.1. It can be seen that the SVM classifier with linear and GHI kernels applied on raw voxel intensity features slightly outperformed the methods applied following the histogram of textons approach, although much worse performances were achieved for the RBF kernel applied on VI features. Moreover, as for the histogram of textons approach, the RBF and GHI kernels were able to achieve similar performances, and, contrary to the theoretical knowledge on classification using SVM, much better than that of the linear kernel. In theory, the increase in the number of features should enhance the ability of the linear SVM to learn a decision boundary in the input space, which, not being confirmed in this case, suggests that the increase in the number of textons considered in the dictionary might not have resulted in an improved predictive power of the input features, so that a

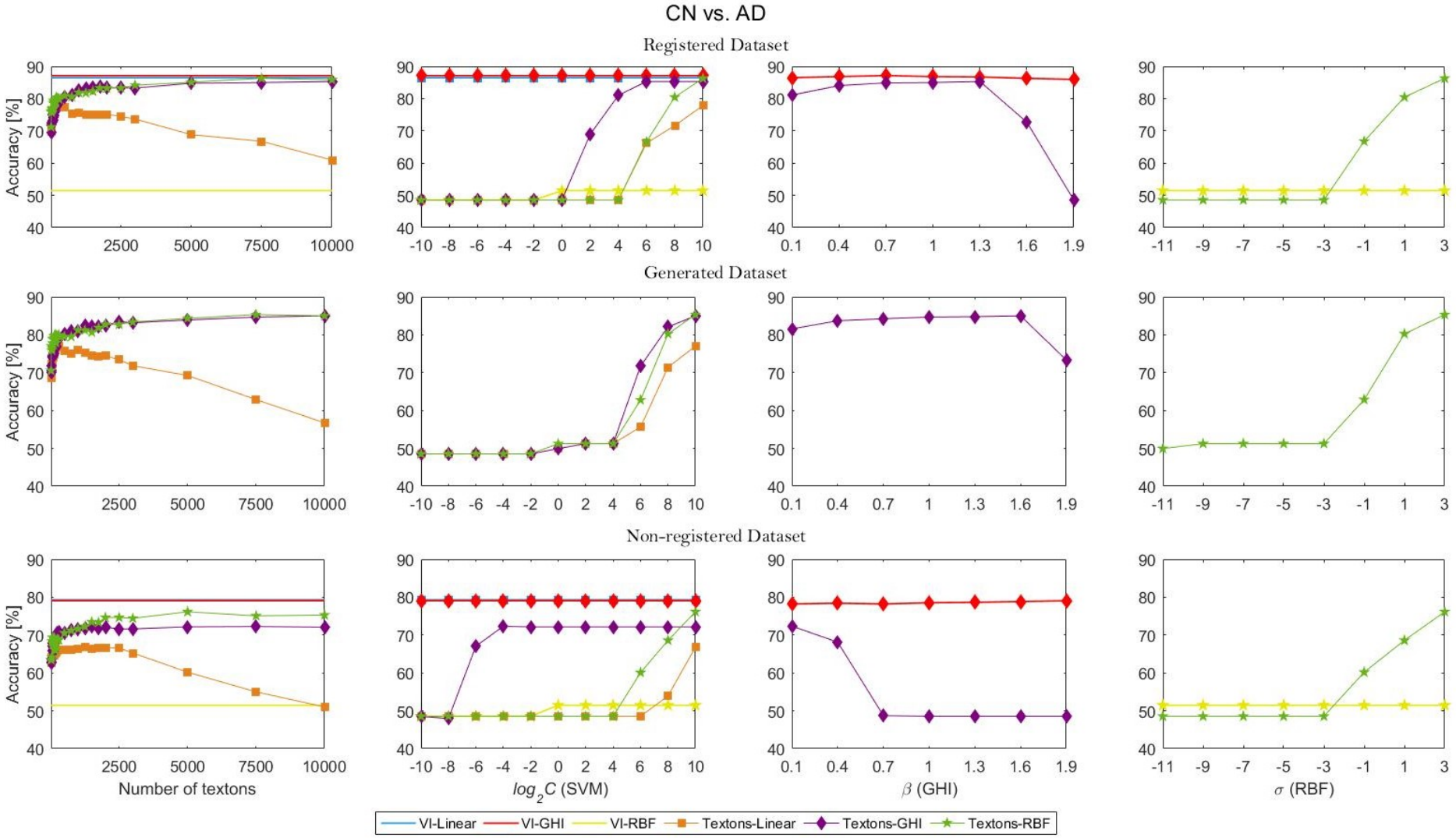


Figure 5.1: Representation of the nested cross-validation procedure for CN vs. AD, using features extracted from the whole brain images.

smaller number of textons should be considered when using the linear kernel (namely below 2500, as suggested by Figure 5.1). On the other hand, the linear kernel performs well once applied to the raw voxel intensity features, which consist of a much larger number of features, and indeed its performance is similar to that of the GHI kernel, confirming the initial hypothesis that when the number of features is extremely large it usually suffices to consider the linear SVM algorithm.

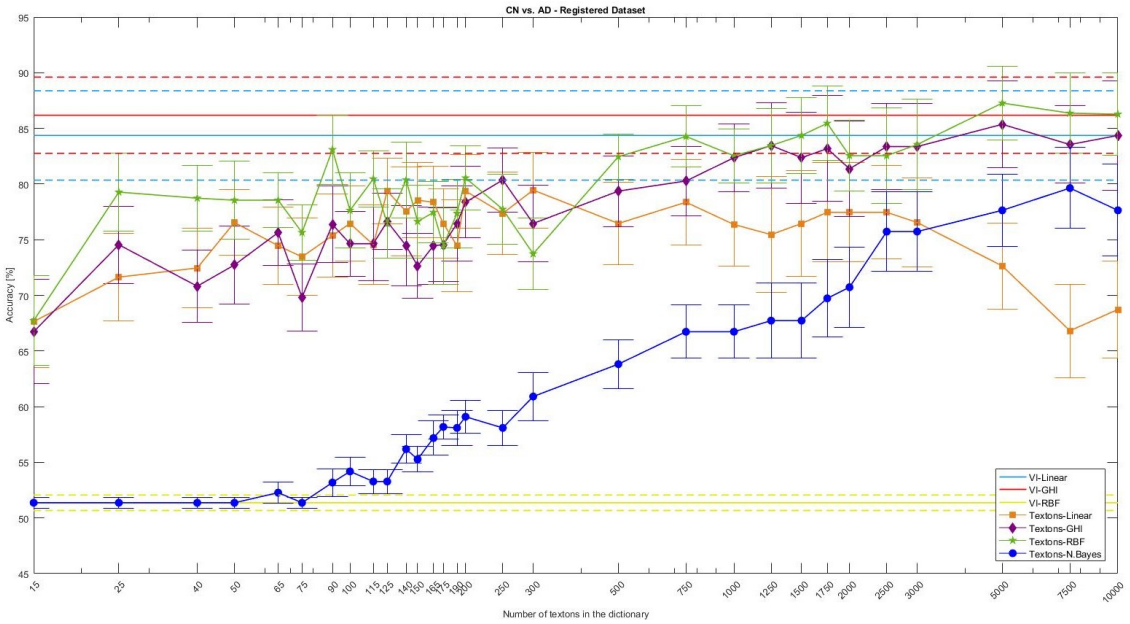
Referring to the penalty parameter C considered in SVM, which trades off correct classification of training examples against maximization of the decision function's margin, a lower value of this variable is thus equivalent to increasing the regularization term, so as to avoid overfitting. From Figure 5.1, it can be observed that in general, and much more evidently for the histogram of textons approach, the classification accuracy increases with C until it stabilizes at the optimal value (within the range considered for the grid-search procedure), which is reached earlier for the GHI kernel than for the remainder, indicating that the SVM algorithm with this kernel might be less prone to overfitting.

Regarding the optimization of the power mapping hyperparameter, β, introduced in the generalized histogram intersection (GHI) kernel, it can be seen that the classification accuracy degrades after the optimal value (again within the range considered for the grid-search procedure) is reached, which occurs at around 1.3 and 1.6 for the registered and non-registered datasets (hence close to the simpler version of using the input histogram features to the power of 1), respectively, and at the lower value of 0.1 for the non-registered one (thus corresponding to scaling down the input features). Using the raw voxel intensities, on the other hand, for the first and latter datasets, it can be seen that the classification accuracy is close to independent from the choice of β.

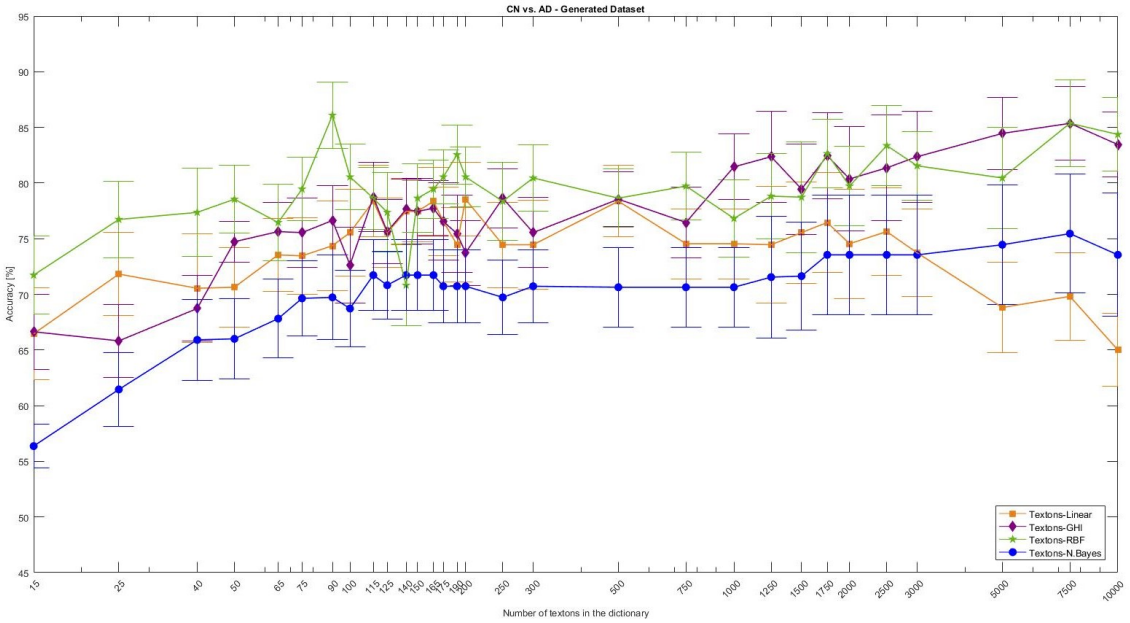
Finally, concerning the parameter σ for the RBF kernel, it concerns the radius of influence of the support vectors, so that lower values thus correspond to, in the limit, using only the support vectors, while much higher values would imply that the whole training set would be considered under its influence, so that the model wouldn't be able to capture the complexity of the data. There is, thus, a trade-off associated to the choice of σ. In this case, while the classification accuracy using the raw voxel intensities as features is close to unaffected by the choice of this parameter, it can be observed that the RBF SVM algorithm fed with the histograms of textons performs better when higher values of σ are considered (namely close to 3, regarding the range used for the grid-search procedure).

Following the nested cross-validation procedure, the final model with the selected hyperparameters was thus applied on the respective test set (for each dataset), leading to the final diagnostic accuracy results for this binary task that are depicted in Figure 5.2. In general, these results, throughout the different datasets, follow the same trend as observed for the previous step, such that the best performances were achieved by the GHI and RBF kernels for the texton-based approach and GHI and linear kernels for the VI-based one. Moreover, the performance of these classifiers didn't degrade much once applied in the independent test set, so that, as desired, no significant overfitting was encountered. It is also important to highligh that while the influence of the number of textons was studied during nested cross-validation, it was not considered as a hyperparameter to optimize, as multiple textons were once again considered in the test set, and the same reasoning applies to the choice of kernel to use in the SVM algorithm.

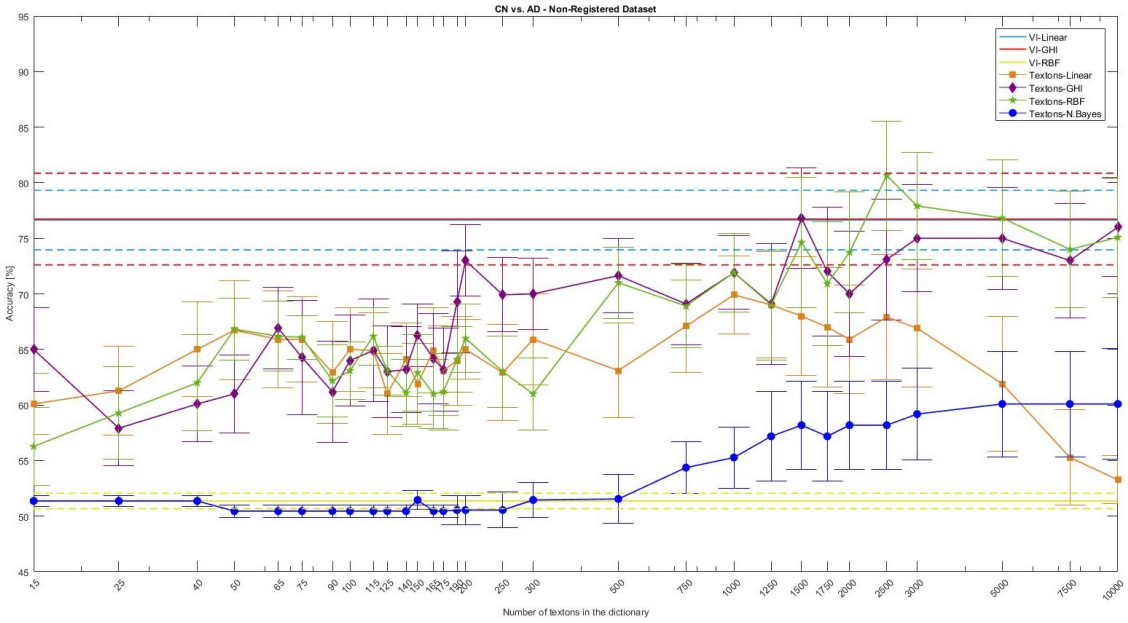
In particular, when using the raw voxel intensities as features, the highest mean diagnostic accuracy


CN vs. AD - Registered Dataset
Accuracy [%]
Number of textons in the dictionary
VI-Linear
VI-GHI
VI-RBF
Textons-Linear
Textons-GHI
Textons-RBF
Textons-N.Bayes

(a)


CN vs. AD - Generated Dataset
Accuracy [%]
Number of textons in the dictionary
Textons-Linear
Textons-GHI
Textons-RBF
Textons-N.Bayes

(b)



(c)

Figure 5.2: Comparison of the results obtained for CN vs. AD, using features extracted from the whole brain, for varying numbers of textons. For both (a) registered, (b) generated and (c) non-registered datasets, mean accuracies are presented, as well as the two standard error of the mean interval (for the texton-based approach), and the upper and lower accuracy bounds (for the VI-based approach).

achieved was around 86.18% for the registered dataset, and 76.73% for the non-registered one, both using an SVM with the GHI kernel, while a much worse performance was observed when using the RBF kernel. A better performance was achieved, in either case, using the texton-based approach, particularly attaining the highest mean accuracy value of 87.27% for the registered dataset (using the RBF kernel, with 5000 textons), 86.09% for the generated one (using the RBF kernel, with 90 textons), and 80.63% for the non-registered dataset (using the RBF kernel, with 2500 textons). As expected, the texton-based approach thus proved to be more robust to skipping of the image registration step, as the use of feature vectors of the voxel intensity within the brain mask should be applied in the case where each particular feature refers to the same anatomical position across all subjects, which is not the case in non-registered datasets. The texton-based approach, on the other hand, can prove to be more adequate for this application as the model extracted for each subject, fed into the classifier, is a histogram of textons, hence disregarding the information of the anatomical structure from where each individual feature originated. Moreover, the fact that in the generated dataset the highest diagnostic accuracy reached was very similar to that of the registered dataset and better than that of the non-registered one, considering that neither this or the latter undergone image registration, might be due to the dimension of the images being much higher for the (real) non-registered dataset, so that an increased number of textons could be required to enhance the classifier's performance, as the task of separating the two classes thus becomes more complicated.

The naive Bayes algorithm was also applied for this final testing step, despite not being considered for the nested cross-validation as it didn't require hyperparameter optimization. It can be observed in

Figure 5.2 that the SVM algorithm outperformed naive Bayes in all the different datasets, and which is particularly evident for the non-registered case. Nonetheless, the much higher complexity and computational cost associated to training the SVM, which is practically negligible for naive Bayes, lead the latter to remain being considered a good alternative, as it can still perform well and is again robust to skipping the image registration step.

5.1.2 Feature extraction from patches of the brain

As mentioned in 4.2, regarding classification using features extracted from patches of the brain, three approaches were considered, namely random patch selection, selection of patches containing the most discriminative textons, and patches within defined regions of interest. In contrast to the previous method, these approaches require the feature vectors to be fed into the classifiers not to be raw voxel intensities as these would depend on the order on which the features are fed and consequently on the anatomical position from which the patches are drawn, hence not being robust to skipping image registration. Thus, the features considered for this section are either histograms of textons or feature representations learned from stacked sparse autoencoders, as introduced in 4.2.

5.1.2.1 Random patch selection

Regarding the classification problem using randomly selected patches of the brain, with a fixed value of 1250 textons in the dictionary, once again all three datasets were tested following, in the case of the SVM algorithm, nested cross-validation for hyperparameter optimization, and the results obtained are presented in Table 5.1. It should be noted that, even though in this method the patches used for classification or training differ between subjects, the classifiers can still perform well, which might be due to the fact that a sufficient number of patches is selected such that there is a high probability of drawing patches within regions that should be relevant for the diagnosis. Using naive Bayes, similar performances could be reported to that for the SVM algorithm, which again highlights the advantage of using this method, for which the computational cost is extremely smaller. The fact that this patch selection procedure is completely random can also explain why better classification results could be reached in the generated dataset than in the registered one, so that this method can be considered to be robust to skipping the image registration step, although its results might be far from optimal and this method should be used as a benchmark for other approaches to outperform.

Table 5.1: Diagnostic accuracy for CN vs. AD, for each dataset, using different classification algorithms on randomly selected patches. Format: Mean (Standard Error of the Mean) [%].

	Registered [%]	Generated [%]	Non-registered [%]
SVM - Linear	79.45 (5.56)	81.27 (4.84)	70.00 (4.92)
SVM - GHI	79.55 (3.76)	80.55 (5.27)	69.00 (4.95)
SVM - RBF	77.64 (4.64)	78.55 (4.16)	69.18 (5.75)
N. Bayes	68.64 (5.71)	73.55 (5.36)	66.82 (4.96)

5.1.2.2 Patches containing discriminative textons

Regarding the selection of patches containing discriminative textons, again the number of textons in the dictionary was fixed at 1250 and the three datasets were considered. Regarding nested cross-validation, once again a similar reasoning considering the parameter tuning can be applied as in the previous methods and the SVM algorithm with GHI kernel still outperforms the remainder in both the registered and generated datasets, while the RBF kernel could lead to better results in the non-registered one, as displayed in Figure 5.3. Concerning the latter dataset, it can also be observed that the diagnostic accuracy is close to invariant to the choice of β in the GHI kernel. As mentioned in 4.2, for this strategy, the number of discriminative textons, computed using the mutual information ranking, from which any should be present in a patch for it to be selected, was also fine tuned through nested cross-validation (not presented in Figure 5.3), varying between 1, 5, 10 or 15 textons, although only slight deviations were observed.

The results drawn from the testing stage, following model selection, are then presented in Table 5.2. In this case, the naive Bayes algorithm was also tested but much poorer performances were attained, so that its results are not presented. Even though the performance of the model in each respective dataset is reasonable, it doesn't outperform the results obtained with the previous seed for random patch selection, so that further methods should be considered and explored. In general, however, in theory this method should attain similar performances regardless of the image registration step, as it does not require any knowledge on the anatomical position from which the patches are drawn and there is a significant overlap between these.

An example of the selected patches containing discriminative textons (in this particular case, resulting from setting the maximum number to 15, as tuned by nested cross-validation, thus corresponding to the larger set of patches considered for classification) for a subject of each class on this binary classification problem, regarding the registered dataset, is presented in Figure 5.4. It can be seen that the set of multiple selected patches for either subject covers the vast majority of the brain, so that the regions identified to be discriminative are found to be similar. An exception to this consists of the posterior cingulate and precuneus, which is still included in the reported regions of interest for diagnosis of Alzheimer's disease, as depicted in Figure 4.13, again corroborating that the texture analysis of this brain region might provide relevant information for diagnosis.

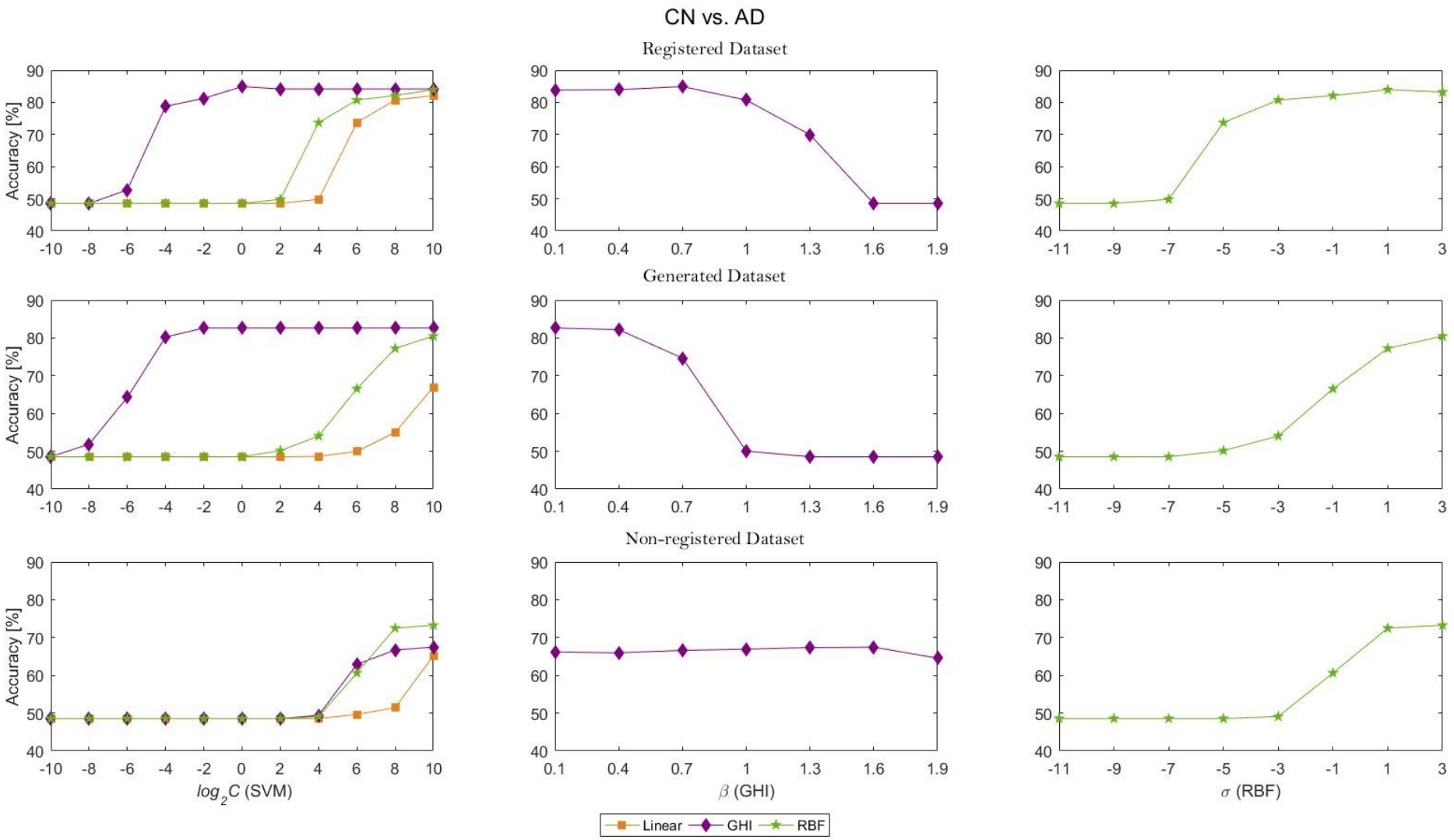


Figure 5.3: Representation of the nested cross-validation procedure for CN vs. AD, using patches containing the most discriminative textons.

Table 5.2: Diagnostic accuracy for CN vs. AD, for each dataset, using different classification algorithms on patches containing the most discriminative textons. Format: Mean (Standard Error of the Mean) [%].

	Registered [%]	Generated [%]	Non-registered [%]
SVM - Linear	77.45 (5.29)	67.00 (5.15)	66.82 (5.79)
SVM - GHI	75.36 (5.22)	70.91 (4.91)	65.09 (5.25)
SVM - RBF	72.55 (5.11)	71.91 (3.36)	68.18 (5.29)

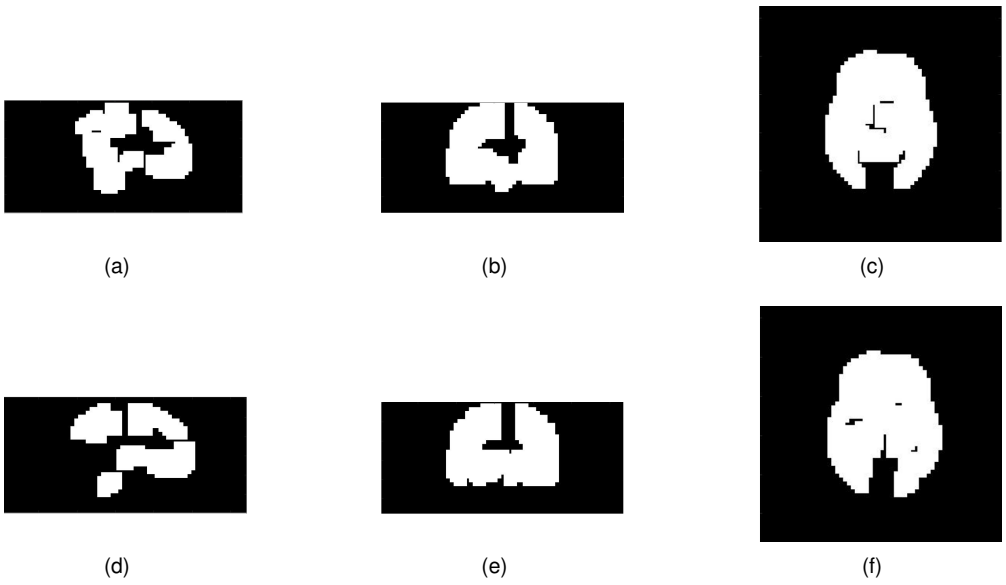


Figure 5.4: Sample images for each class illustrating the selection of patches containing the most discriminative textons. Figures (a) to (c) and (d) to (f) represent sagittal, coronal and axial sections of the brain of a subject in the CN and in the AD class, respectively.

5.1.2.3 Patches within ROIs

Regarding the classification problem using patches within the reported ROIs (depicted in Figure 4.13), as mentioned in 4.2 it was assumed that image registration would still have to be performed, but only on the training stage. The patches were then classified, at the testing stage, according to the previously labeled regions, so as to be used, or not, for the final diagnosis of the test subject. The results of each of these stages (patch selection and final image classification), applied on both the registered and non-registered datasets retrieved from ADNI, are presented in the following subsections.

5.1.2.3.1 Patch selection

Concerning patch selection, a deep learning strategy was employed to learn feature representations of the input patches, namely a stacked sparse autoencoder, followed by an output softmax classifier layer, as mentioned in 4.2, and the results are presented in Table 5.3. It can be observed that the accuracy in patch selection using this method was very high regarding the registered dataset, while worse performances were achieved for the non-registered one, as would be expected being the latter a more difficult classification problem. Moreover, since this step is performed and required prior to the

final binary classification between CN and AD, the error obtained regarding patch selection propagates to that final step, as the final diagnosis will be performed using patches within regions that might not be affected by the disease, thus possibly reducing the sensitivity of the method.

An example of the selected patches using this method, regarding the registered dataset, is presented in Figure 5.5. It can be seen that the set of multiple selected patches for either subject covers the vast majority of the reported ROIs (with the exception, for instance, of the right lateral temporal, not identified in the sample subject of the AD class), as well as other regions that might not be affected by the disease.

Table 5.3: Patch selection accuracy for CN vs. AD, for both the registered and non-registered datasets, considering previously identified ROIs. Format: Mean (Standard Error of the Mean) [%].

	Registered [%]	Non-registered [%]
SSAE+Softmax	91.18 (0.13)	79.74 (0.23)

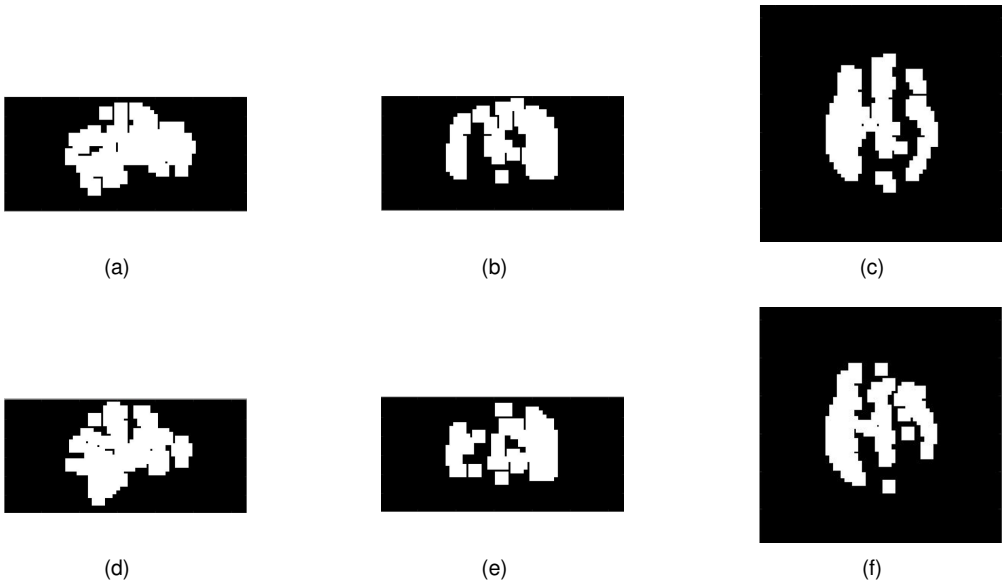


Figure 5.5: Sample images for each class illustrating the selection of patches within ROIs. Figures (a) to (c) and (d) to (f) represent sagittal, coronal and axial sections of the brain of a subject in the CN and in the AD class, respectively.

5.1.2.3.2 Image Classification

The patches previously selected were then fed as input to the final classification algorithm between CN and AD. Several methods were explored in this step, namely using again an SSAE and softmax classifier at the output layer where each patch was first classified in either CN or AD through supervised learning, and followed by majority voting to obtain the final diagnosis. Alternatively, the histogram of textons on the reunion of all selected patches was also computed (again using 1250 textons in the dictionary) and fed into an SVM or naive Bayes classification algorithm, where in the former the three kernels previously mentioned were again tested (namely GHI, RBF and linear). Another method that was proposed consisted of again attaining the final diagnosis through majority voting in combination with the

histogram of textons computed for each patch. Nonetheless, due to the extremely high computational cost associated to the nested cross-validation procedure for optimization of the hyperparameters of the SVM algorithm, and due to the fact that the final diagnostic accuracy could largely depend on these parameters such that fixing specific ones could also be misleading, and that naive Bayes could also not perform well in this task, as initially tested, these results were excluded from this discussion and hence not presented in Table 5.4.

From the different methods explored, it can be observed SSAE followed by softmax performed very well on the registered dataset, reaching a diagnostic accuracy close to the state of the art (close to 90%), while the texton-based approach could also lead to good results, namely using the linear kernel. It is crucial to highlight that, due to the high computational cost associated to training the full network, an option was made to perform the analysis regarding the non-registered dataset using only 2 folds (as indicated by * in Table 5.4) for the cross-validation procedure, unlike the previous case of 10 considered in the remaining analyses, so that the results here displayed for the latter dataset, despite being quite satisfactory, could eventually be positively biased, hence becoming inconclusive.

Table 5.4: Patch selection accuracy for CN vs. AD, for both the registered and non-registered datasets, considering previously selected patches within ROIs. Format: Mean (Standard Error of the Mean) [%].

	Registered [%]	Non-registered* [%]
SVM - Linear (Reunion)	81.45 (3.50)	76.36 (3.64)
SVM - GHI (Reunion)	76.36 (4.84)	76.36 (3.64)
SVM - RBF (Reunion)	77.36 (4.27)	76.36 (3.64)
N. Bayes (Reunion)	62.62 (4.68)	71.82 (3.66)
SSAE+Softmax (Maj. Voting)	89.09 (4.07)	85.45 (5.45)

5.2 Cognitively normal vs. mild cognitive impairment

5.2.1 Feature Extraction From The Whole Brain

As in the previous binary classification task, in the CN vs. MCI problem, the classification algorithms applied on features extracted from the whole brain were also tested on all three of the datasets introduced in 4.2.1. The parameters associated with the different classifiers were again tuned inside a nested cross-validation procedure, according to 4.2, as shown in Figure 5.6. It can be seen that the SVM classifier with both linear, RBF and GHI kernels applied on the texton-based approach slightly outperformed the methods applied on raw voxel intensity features (again only tested on the two real datasets retrieved from ADNI, and for which the results using the RBF kernel are not represented due to its very poor performance). Moreover, contrary to what occurred for the previous binary classification task, here, as expected, the performance of the SVM with linear kernel doesn't degrade with the increase in the number of features and, in general, better diagnostic results can be attained with larger numbers of textons. Moreover, better results in the nested cross-validation procedure are reached when a higher penalty parameter is considered, so that smaller regularization terms should be introduced in the final classifier.

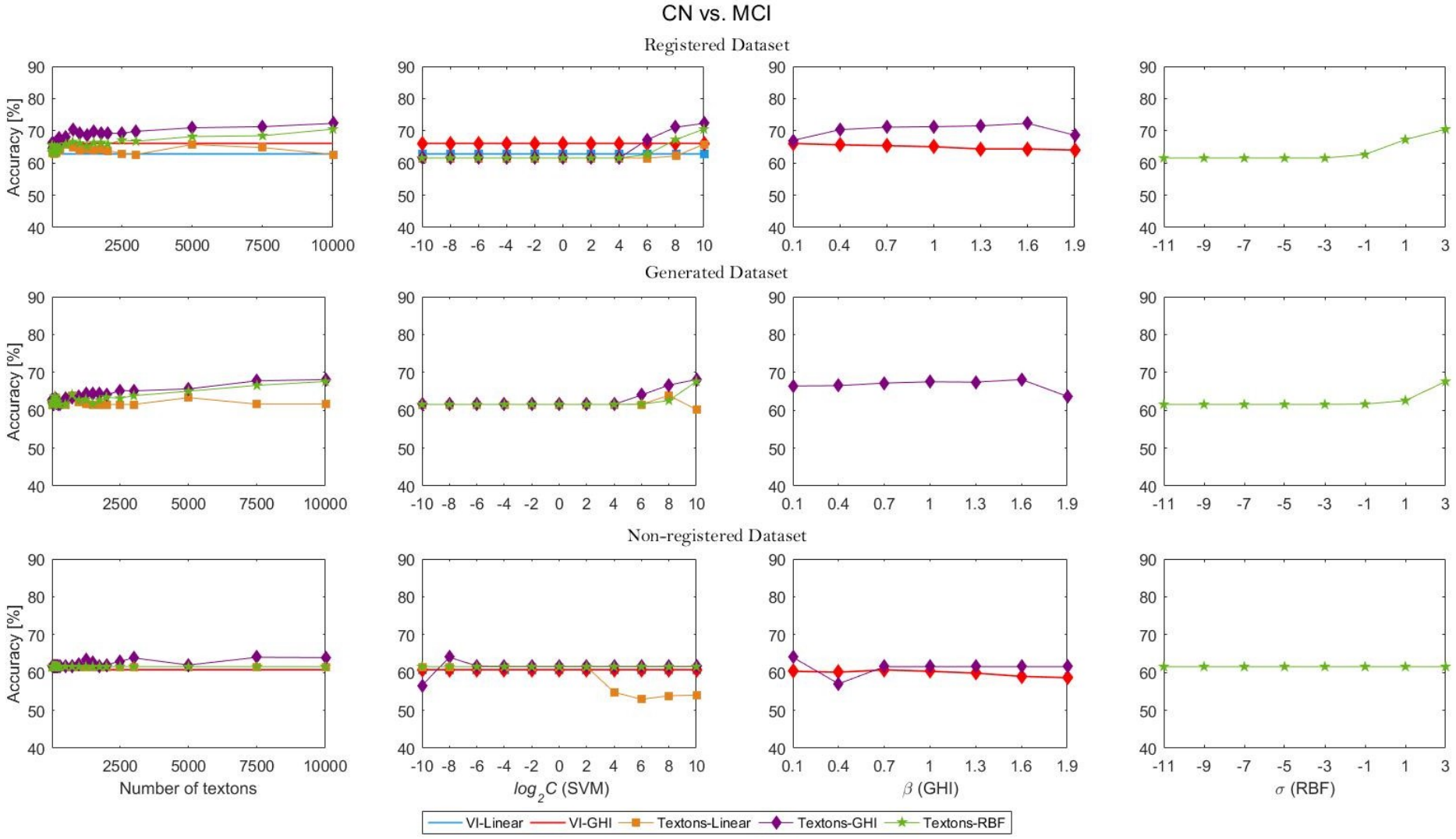
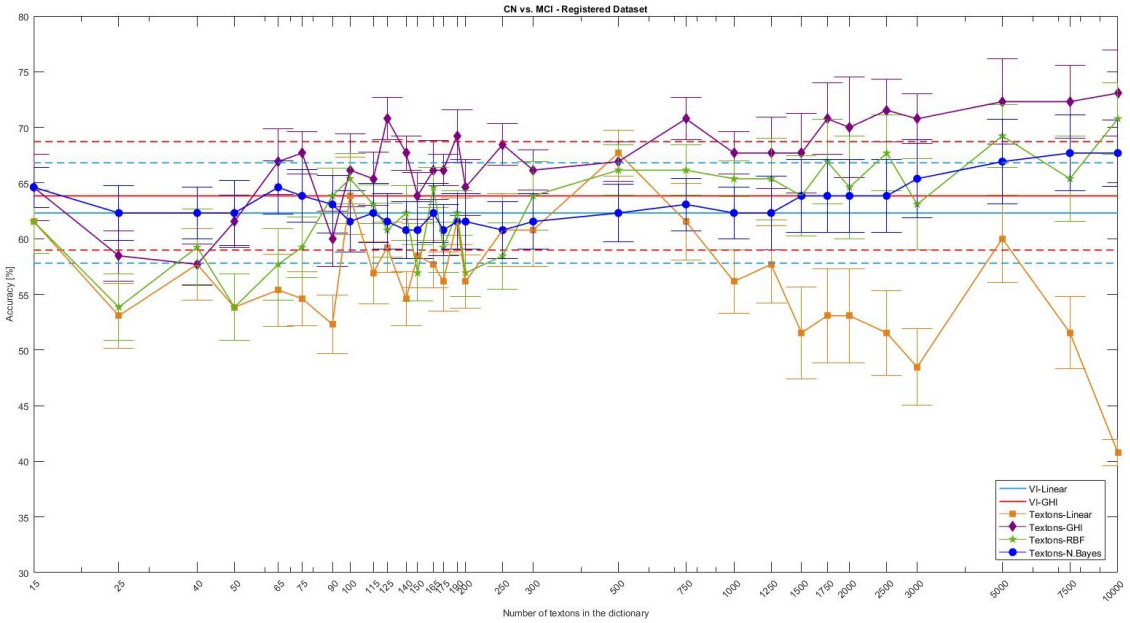


Figure 5.6: Representation of the nested cross-validation procedure for CN vs. MCI, using features extracted from the whole brain images.

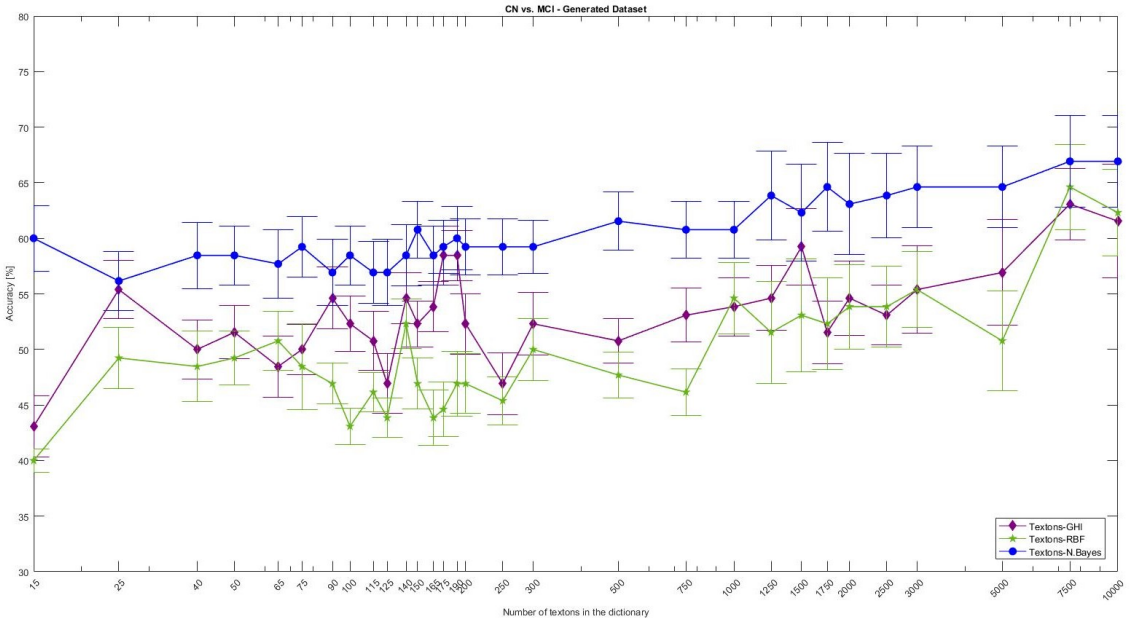
Finally, the oscillations in accuracy through the range of the values of β and σ considered are again almost negligible.

Again following this procedure for hyperparameter optimization, the selected model was applied on the respective test set (for each dataset), leading to the final diagnostic accuracy results for this binary task that are depicted in Figure 5.7. In general, these results, throughout the different datasets, follow the same trend as observed for the previous step, such that the best performances were achieved by the GHI and RBF kernels for the texton-based approach. The results from applying the GHI kernel on raw voxel intensity features in the non-registered dataset, and the linear kernel for both non-registered ones, as well as for the RBF kernel in the final dataset, are not depicted due to the fact that these attained very poor performances. Again, multiple textons were considered on the test set, not being defined as a hyperparameter to optimize through nested cross-validation.

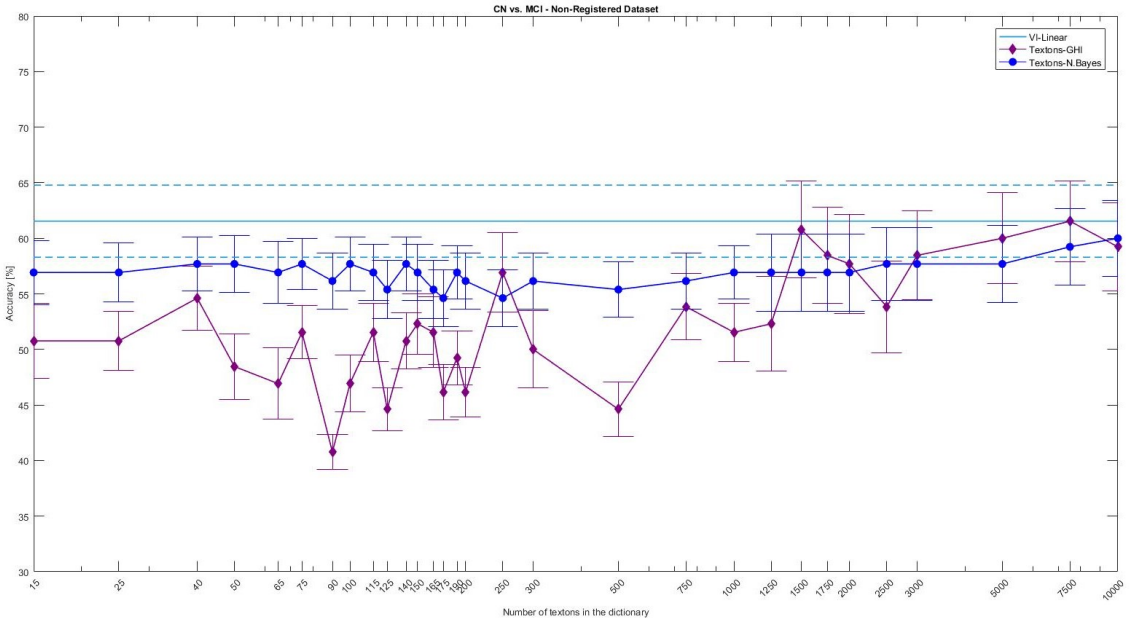
Concerning the use of raw voxel intensities as features, the highest mean diagnostic accuracy achieved was around 63.85% for the registered dataset, and 61.54% for the non-registered one, both using the GHI kernel. A better performance was achieved, in either case, using the texton-based approach, particularly attaining the highest mean accuracy value of 73.08% for the former (using 10000 textons) and 61.54% for the latter (using 7500 textons), while for the generated dataset, the naive Bayes algorithm (again also tested without requiring parameter tuning) reached the highest mean accuracy value, particularly of 66.92%.



(a)



(b)



(c)

Figure 5.7: Comparison of the results obtained for CN vs. MCI, using features extracted from the whole brain, for varying numbers of textons. For both (a) registered, (b) generated and (c) non-registered datasets, mean accuracies are presented, as well as the two standard error of the mean interval (for the texton-based approach), and the upper and lower accuracy bounds (for the VI-based approach).

Similar to the previous task, presented in 5.1.1, it can be observed that the performance of the majority of the different models would improve when considering larger numbers of textons in the dictionary, so that it could be hypothesized that better diagnostic accuracies could have been attained in case the number of textons in the dictionary was extended beyond 10000; this should also be particularly true

in the non-registered dataset, as the dimensions of these images are much larger than that for the registered ones. However, this would bring the disadvantage of requiring a higher computational cost, resulting in a trade-off between the two.

5.2.2 Feature extraction from patches of the brain

5.2.2.1 Random patch selection

Regarding the classification problem using randomly selected patches of the brain, with a fixed value of 1250 textons in the dictionary, once again all three datasets were tested following, in the case of the SVM algorithm, nested cross-validation for hyperparameter optimization. It can be observed that the results from applying the naive Bayes algorithm are comparable to the former, while significantly reducing the complexity and computational cost of the problem, although all models exhibit very poor performances (close to a random procedure for binary classification).

Table 5.5: Diagnostic accuracy for CN vs. MCI, for each dataset, using different classification algorithms on randomly selected patches. Format: Mean (Standard Error of the Mean) [%].

	Registered [%]	Generated [%]	Non-registered [%]
SVM - Linear	54.62 (4.36)	43.08 (3.48)	39.23 (0.77)
SVM - GHI	66.92 (3.81)	55.38 (2.99)	56.92 (3.84)
SVM - RBF	64.62 (4.17)	53.08 (5.06)	42.31 (3.85)
N. Bayes	62.31 (3.13)	62.31 (5.19)	59.23 (3.45)

5.2.2.2 Patches containing discriminative textons

Regarding the selection of patches containing discriminative textons, again the number of textons in the dictionary was fixed at 1250 and the three datasets were considered. The results drawn from the testing stage, following model selection through nested cross-validation, are then presented in Table 5.6, where once again the naive Bayes algorithm was also tested but much poorer performances were attained, so that its results are not presented. As occurred in the previous binary classification task, the applied algorithms for this approach do not outperform the results obtained with the previous seed for random patch selection, so that further techniques should be considered and explored.
Nonetheless, and as mentioned in 5.1.2.2, this method should be able to exhibit similar performances regardless of the image registration step, as it does not require any knowledge on the anatomical position from which the patches are drawn and there is a significant overlap between these.

An example of the selected patches containing discriminative textons (in this case, resulting from setting the maximum number to 15, as tuned by nested cross-validation) for a subject of each class on this binary classification problem, regarding the registered dataset, is presented in Figure 5.8. It can be observed that the regions found to be discriminative through this method approximate the previously reported ROIs (reported for the diagnosis of Alzheimer's disease), illustrated in Figure 4.13, namely considering the inferior anterior cingulate, which again corroborates that the texture analysis of these regions might provide relevant information for diagnosis, including for the case of MCI.

Table 5.6: Diagnostic accuracy for CN vs. MCI, for each dataset, using different classification algorithms on patches containing the most discriminative textons. Format: Mean (Standard Error of the Mean) [%].

	Registered [%]	Generated [%]	Non-registered [%]
SVM - Linear	56.15 (3.45)	49.23 (3.66)	47.69 (3.40)
SVM - GHI	66.92 (3.04)	60.77 (4.51)	62.31 (3.71)
SVM - RBF	57.69 (3.29)	49.23 (4.17)	44.62 (8.01)

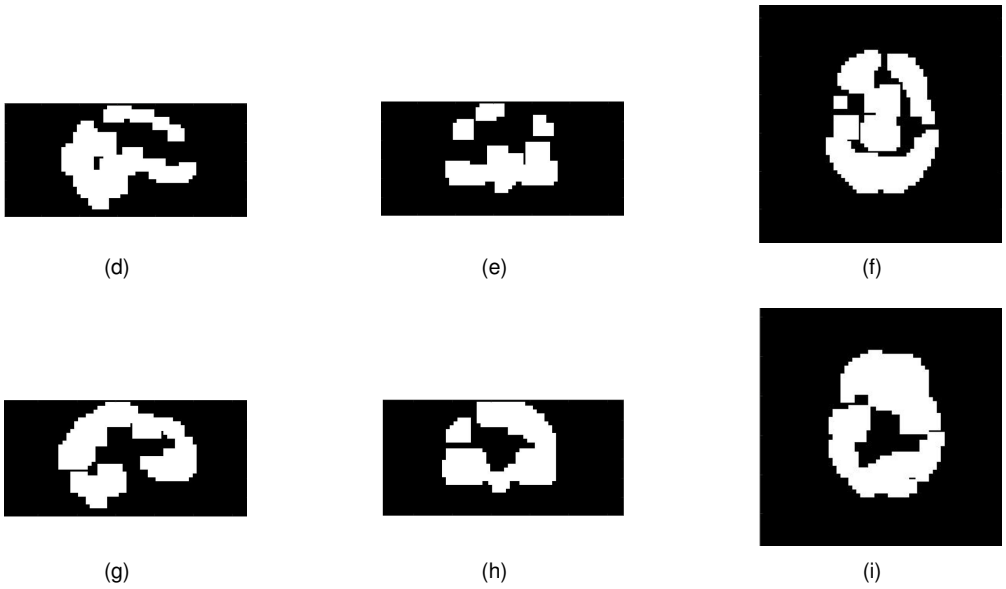


Figure 5.8: Sample images for each class illustrating the selection of patches containing the most discriminative textons. Figures (a) to (c) and (d) to (f) represent sagittal, coronal and axial sections of the brain of a subject in the CN and in the MCI class, respectively.

5.2.2.3 Patches within ROIs

As in the previous binary classification problem, for this approach two steps were performed, namely patch selection and the final image classification, for both the registered and non-registered datasets retrieved from ADNI, as presented in the following subsections.

5.2.2.3.1 Patch selection

The attained results in this step are presented in Table 5.7. It can be observed that while the SSAE with softmax classifier performed very well in the registered dataset, worse performances were achieved for the non-registered one, being theoretically a more complicated classification problem.

An example of the selected patches using this method, regarding the registered dataset, is presented in Figure 5.9, where it can be seen that the set of multiple selected patches for either subject covers the vast majority of the reported ROIs, as would be desired.

Table 5.7: Patch selection accuracy for CN vs. MCI, for both the registered and non-registered datasets, considering previously identified ROIs. Format: Mean (Standard Error of the Mean) [%].

	Registered [%]	Non-registered [%]
SSAE+Softmax	91.55 (0.12)	79.92 (0.12)

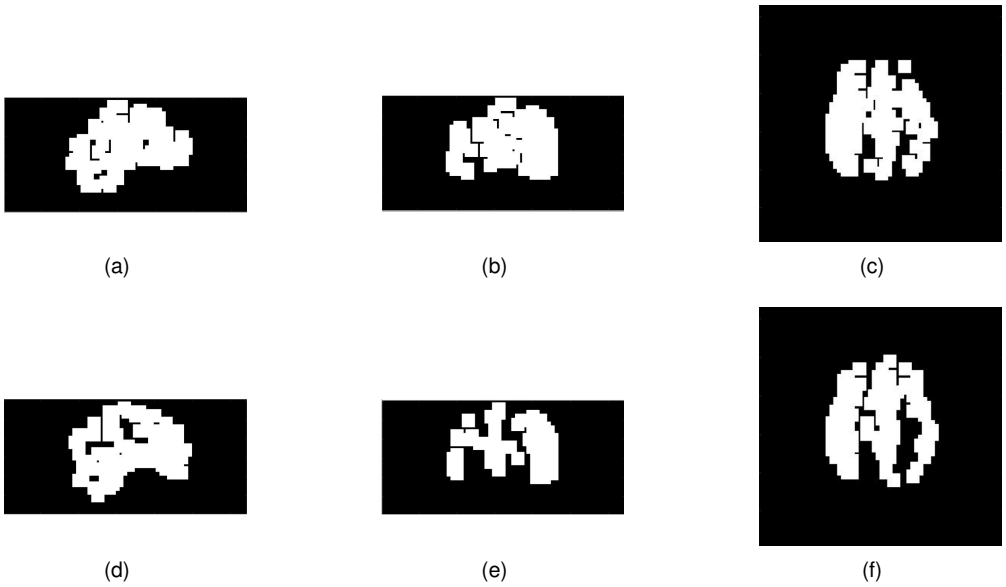


Figure 5.9: Sample images for each class illustrating the selection of patches within ROIs. Figures (a) to (c) and (d) to (f) represent sagittal, coronal and axial sections of the brain of a subject in the CN and in the MCI class, respectively.

5.2.2.3.2 Image classification

As in the CN vs. AD classification task, the selected patches were then fed as input to the final classification algorithm between CN and MCI. Several methods were explored in this step, namely using again an SSAE and softmax classifier followed by majority voting, or alternatively the histogram of textons on the reunion of all selected patches was also computed (again using 1250 textons in the dictionary) and fed into an SVM or naive Bayes classification algorithm, where in the former the three kernels previously mentioned were again tested (namely GHI, RBF and linear).Again, due to the inherent high computational cost, the texton-based approach was not considered in combination with majority voting.

From the different methods explored, it can be observed that the method which performed betterr consisted on the linear SVM applied to the reunion of selected patches for the registered dataset, while the SSAE followed by softmax and majority voting led to the highest classification accuracy on the non-registered dataset. Once again, it should be higlighted that only 2 folds (as indicated by * in Table 5.4) were used for the cross-validation procedure on the non-registered dataset, so that its results could eventually be positively biased, hence becoming inconclusive. Moreover, the propagation of error in patch selection might have led to worse performances, particularly in the non-registered dataset.

Table 5.8: Patch selection accuracy for CN vs. MCI, for both the registered and non-registered datasets, considering previously selected patches within ROIs. Format: Mean (Standard Error of the Mean) [%].

	Registered [%]	Non-registered* [%]
SVM - Linear (Reunion)	73.85 (4.47)	42.31 (3.85)
SVM - GHI (Reunion)	74.62 (3.98)	61.54 (3.44)
SVM - RBF (Reunion)	77.69 (4.51)	42.31 (3.85)
N. Bayes (Reunion)	65.38 (2.63)	46.15 (0)
SSAE+Softmax (Maj. Voting)	61.54 (0)	61.54 (0)

5.3 Summary

The results obtained for the CN vs. AD and CN vs. MCI binary classification problems, using the different methods explored, are summarized in Table 5.9 and Table 5.10, respectively.

Regarding the CN vs. AD problem, the best accuracy results obtained were 89.09%, 86.09% and 80.63%, respectively for the registered, generated and non-registered datasets, the former using patch selection and majority voting by means of SSAE and the softmax classifier, and the remainder using textons extracted from the whole brain images, which proved to be robust to registration errors. High sensitivity values are also reported for the three datasets, indicating that a large percentage of patients with Alzheimer's disease can thus be correctly classified, as desired. Regarding specificity, high values are also reported for all datasets, indicating a low ratio of false positives (meaning, of a patient being classified as belonging to the AD class, while being in fact cognitively normal).

Concerning the CN vs. MCI problem, being a much more complicated classification task, the attained results in terms of both accuracy, sensitivity and specificity were naturally lower. The best accuracy results obtained were 77.69%, 66.92% and 62.31%, respectively for the registered, generated and non-registered datasets, the former using patch selection and further classification using histogram of textons on the reunion of all patches, while the remainder using textons extracted from the whole brain images and patches containing discriminative textons, respectively. For the former method, a value of 100% was reported for the sensitivity in the registered dataset, meaning that all subjects with mild cognitive impairment could be correctly classified. However, this was obtained at the cost of a 50% specificity, meaning that all subjects (even cognitively normal) were indeed classified as having MCI, so that the method's performance was not particularly satisfactory. Specificity values close to 60% were also reported for the majority of the methods explored in this classification problem, being considered reasonable. Nonetheless, considering that it might be more relevant to accurately predict the diagnosis for subjects having MCI or AD, the results obtained for both methods might in this sense be considered satisfactory, as the majority of these could perform better considering sensitivity than specificity.

In the discussion above, the results from approach considering patches within ROIs were disregarded for the non-registered datasets in both CN vs. AD and CN vs. MCI since, as previously described, these reuslts are not directly comparable to that of the remaining methods, as the cross-validation procedure considered only 2 folds instead of the usual 10, which could have largely contributed with a positive bias

towards these results.

Table 5.9: Summary of the results obtained for CN vs. AD, for all datasets and methods explored. The best results are presented in bold. The * indicates that 2 folds (and not 10) were used for cross-validation.

	Registered [%]			Generated [%]			Non-registered [%]		
	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE
Whole brain - VI	86.18	88	84	-	-	-	76.73	**77.33**	76
Whole brain - Textons	87.27	86.67	88	**86.09**	**86**	**86**	**80.63**	73.67	**88**
Random patches	79.55	81	78	81.27	76.67	86	70.00	62.67	78
Discriminative textons	77.45	72.67	82	71.91	71.67	72	68.18	66.33	70
Patch Selection + Softmax	**89.09**	**90**	**88**	-	-	-	**85.45***	**100***	70*
Patch Selection + Reunion	81.45	82	80	-	-	-	76.36*	63.33*	**90***

Table 5.10: Summary of the results obtained for CN vs. MCI, for all datasets and methods explored. The best results are presented in bold. The * indicates that 2 folds (and not 10) were used for cross-validation.

	Registered [%]			Generated [%]			Non-registered [%]		
	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE
Whole brain - VI	63.85	83.75	32	-	-	-	61.54	70	48
Whole brain - Textons	73.08	81.25	**60**	**66.92**	70	**62**	61.54	67.5	52
Random patches	66.92	73.75	56	62.31	65	58	59.23	60	**58**
Discriminative textons	66.92	71.25	60	60.77	**73.75**	40	**62.31**	**72.5**	46
Patch Selection + Softmax	61.54	**100**	50	-	-	-	61.54*	**100***	50*
Patch Selection + Reunion	**77.69**	90	58	-	-	-	61.54*	87.5*	20*

Chapter 6

Conclusions

This thesis aimed at developing a supervised machine learning tool for computer aided diagnosis of Alzheimer's disease through which the image registration step could be disregarded. Several methods were explored for this purpose, using features extracted from the whole brain, patches or previously labeled regions of interest, which in turn corresponded to the raw voxel intensity values, the extracted histogram of textons or the learned feature representations through the deep learning strategy considered, namely a stacked sparse autoencoder. In order to be able to draw conclusions regarding the robustness of the different methods to registration errors, three datasets were considered, two consisting of real data collections retrieved from the ADNI online database, and the latter being artificially generated through affine transforms applied on the registered dataset.

The different methods proposed were exhaustively studied, with many different models and combinations of feature extraction and transformation followed by the final classification algorithms being tested on the different datasets, which allowed for a clear understanding of the problem that motivated this work. While the main objective of this thesis was reached, many other interesting approaches could have been considered, as discussed in the following subsections. It is also important to highlight that, in the future, the high computational cost associated to the simulations performed when considering several of these methods is an issue that should be adressed and that consists of a limitation of this work, so that, in more efficient methods should be hypothesized.

6.1 Achievements

Regarding the achievements of this work, it can be considered that its major objective was reached, as both the texton-based approach using features extracted from the whole brain images, as well as the learned feature representations using the stacked sparse autoencoder could lead to a good performance of the respective method when being applied in each binary classification problem (although this result was much more evident for the CN vs. AD task), regardless of the application or not of the image registration step.

6.2 Future Work

Concerning possible approaches to be tested in the future, as mentioned throughout this work, it would be quite interesting to study how switching from using the filter bank applied here (namely the 3D extension to the MR8 filter bank, as described) to a set of filters (weights) to be learned by a stacked sparse autoencoder would influence the performance of the models in each classification problem, when using the same texton-based approach. Further studies could also be performed to expand the results presented in this work, namely performing majority voting on the patches selected through SSAE combined with the softmax classifier, or else also considering its application to the whole brain images.

Other deep learning strategies that have attained state of the art performances could also be explored, as presented in Section 3.6, namely combining autoencoders with convolutional neural networks, or else considering deep Boltzmann machines, amongst many other promising techniques.

Bibliography

[1] A. Association et al. 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 14(3): 367–429, 2018.

[2] A. Burns and S. Iliffe. Alzheimer's disease. *BMJ*, 338, 2009. ISSN 0959-8138. doi: 10.1136/bmj.b158. URL https://www.bmj.com/content/338/bmj.b158.

[3] Alzheimer's Association. About the Alzheimer's Association, 2018. URL https://www.alz.org/about. Last access: September 20th, 2018.

[4] L. E. Hebert, J. Weuve, P. A. Scherr, and D. A. Evans. Alzheimer's disease in the United States (2010–2050) estimated using the 2010 census. *Neurology*, 80(19):1778–1783, 2013.

[5] R. Brookmeyer, N. Abdalla, C. H. Kawas, and M. M. Corrada. Forecasting the prevalence of preclinical and clinical Alzheimer's disease in the United States. *Alzheimer's & Dementia*, 14(2):121–129, 2018.

[6] G. B. Frisoni, M. Boccardi, F. Barkhof, K. Blennow, S. Cappa, K. Chiotis, J.-F. Démonet, V. Garibotto, P. Giannakopoulos, A. Gietl, et al. Strategic roadmap for an early diagnosis of Alzheimer's disease based on biomarkers. *The Lancet Neurology*, 16(8):661–676, 2017.

[7] T. Khan. Chapter 2 - Clinical diagnosis of Alzheimer's disease. In T. Khan, editor, *Biomarkers in Alzheimer's Disease*, pages 27 – 48. Academic Press, 2016. ISBN 978-0-12-804832-0. doi: https://doi.org/10.1016/B978-0-12-804832-0.00002-X. URL http://www.sciencedirect.com/science/article/pii/B978012804832000002X.

[8] D. Anchisi, B. Borroni, M. Franceschi, N. Kerrouche, E. Kalbe, B. Beuthien-Beumann, S. Cappa, O. Lenz, S. Ludecke, A. Marcone, et al. Heterogeneity of brain glucose metabolism in mild cognitive impairment and clinical progression to Alzheimer's disease. *Archives of neurology*, 62(11):1728–1733, 2005.

[9] K. A. Johnson, N. C. Fox, R. A. Sperling, and W. E. Klunk. Brain imaging in Alzheimer's disease. *Cold Spring Harbor perspectives in medicine*, page a006213, 2012.

[10] Y. Ou, H. Akbari, M. Bilello, X. Da, and C. Davatzikos. Comparative evaluation of registration algorithms in different brain databases with varying difficulty: results and insights. *IEEE transactions on medical imaging*, 33(10):2039–2065, 2014.

[11] D. P. Shamonin, E. E. Bron, B. P. Lelieveldt, M. Smits, S. Klein, and M. Staring. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. *Frontiers in neuroinformatics*, 7:50, 2014.

[12] P. Morgado, M. Silveira, and D. C. Costa. Texton-based diagnosis of Alzheimer's disease. In *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, pages 1–6. IEEE, 2013.

[13] A. Payan and G. Montana. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *arXiv preprint arXiv:1502.02506*, 2015.

[14] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. *IEEE transactions on medical imaging*, 35(1):119–130, 2016.

[15] J. Neugroschl and S. Wang. Alzheimer's disease: diagnosis and treatment across the spectrum of disease severity. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, 78(4):596–612, 2011.

[16] R. H. Takahashi, T. Nagao, and G. K. Gouras. Plaque formation and the intraneuronal accumulation of β-amyloid in Alzheimer's disease. *Pathology international*, 67(4):185–193, 2017.

[17] D. Harman. Alzheimer's disease pathogenesis. *Annals of the New York Academy of Sciences*, 1067(1):454–460, 2006.

[18] D. J. Selkoe. The molecular pathology of Alzheimer's disease. *Neuron*, 6(4):487–498, 1991.

[19] Y.-L. Chang, M. Bondi, L. McEvoy, C. Fennema-Notestine, D. Salmon, D. Galasko, D. Hagler, A. Dale, A. D. N. Initiative, et al. Global clinical dementia rating of 0.5 in MCI masks variability related to level of function. *Neurology*, 76(7):652–659, 2011.

[20] C. Woolf, M. J. Slavin, B. Draper, F. Thomassen, N. A. Kochan, S. Reppermund, J. D. Crawford, J. N. Trollor, H. Brodaty, and P. S. Sachdev. Can the clinical dementia rating scale identify mild cognitive impairment and predict cognitive and functional decline? *Dementia and geriatric cognitive disorders*, 41(5-6):292–302, 2016.

[21] J. C. Morris. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology*, 1993.

[22] R. Duara, D. A. Loewenstein, M. T. Greig-Custo, A. Raj, W. Barker, E. Potter, E. Schofield, B. Small, J. Schinka, Y. Wu, et al. Diagnosis and staging of mild cognitive impairment, using a modification of the clinical dementia rating scale: the mCDR. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, 25(3):282–289, 2010.

[23] A. Nordberg, J. O. Rinne, A. Kadir, and B. Långström. The use of PET in Alzheimer's disease. *Nature Reviews Neurology*, 6(2):78, 2010.

[24] K. Herholz, S. Carter, and M. Jones. Positron emission tomography imaging in dementia. *The British journal of radiology*, 80(special_issue_2):S160–S167, 2007.

[25] P. Edison, H. Archer, R. Hinz, A. Hammers, N. Pavese, Y. Tai, G. Hotton, D. Cutler, N. Fox, A. Kennedy, et al. Amyloid, hypometabolism, and cognition in Alzheimer's disease: An [^{11}C]-PIB and [^{18}F]-FDG-PET study. *Neurology*, 68(7):501–508, 2007.

[26] T. Kato, Y. Inui, A. Nakamura, and K. Ito. Brain fluorodeoxyglucose (FDG) pet in dementia. *Ageing research reviews*, 30:73–84, 2016.

[27] C. Marcus, E. Mena, and R. M. Subramaniam. Brain PET in the diagnosis of Alzheimer's disease. *Clinical nuclear medicine*, 39(10):e413, 2014.

[28] N. Okamura, R. Harada, S. Furumoto, H. Arai, K. Yanai, and Y. Kudo. Tau PET imaging in alzheimer's disease. *Current neurology and neuroscience reports*, 14(11):500, 2014.

[29] I. Garali, M. Adel, S. Bourennane, and E. Guedj. Region-based brain selection and classification on PET images for Alzheimer's disease computer aided diagnosis. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 1473–1477. IEEE, 2015.

[30] M. Wehenkel, C. Bastin, P. Geurts, and C. Phillips. Computer aided diagnosis system based on random forests for the prognosis of Alzheimer's disease. In *1st HBP Student Conference-Transdisciplinary Research Linking Neuroscience, Brain Medicine and Computer Science*. Frontiers Media SA, 2018.

[31] P. Coupé, S. F. Eskildsen, J. V. Manjón, V. S. Fonov, J. C. Pruessner, M. Allard, D. L. Collins, A. D. N. Initiative, et al. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. *NeuroImage: clinical*, 1(1):141–152, 2012.

[32] M. Liu, D. Zhang, D. Shen, A. D. N. Initiative, et al. Ensemble sparse classification of Alzheimer's disease. *NeuroImage*, 60(2):1106–1116, 2012.

[33] M. Liu, D. Zhang, D. Shen, and A. D. N. Initiative. Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. *Human brain mapping*, 35(4):1305–1319, 2014.

[34] L. Khedher, J. Ramírez, J. M. Górriz, A. Brahim, F. Segovia, A. s Disease Neuroimaging Initiative, et al. Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images. *Neurocomputing*, 151:139–150, 2015.

[35] Y. Zhang, Z. Dong, P. Phillips, S. Wang, G. Ji, J. Yang, and T.-F. Yuan. Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. *Frontiers in Computational Neuroscience*, 9:66, 2015.

[36] M. M. Dessouky, M. A. Elrashidy, T. E. Taha, and H. M. Abdelkader. Computer-aided diagnosis system for Alzheimer's disease using different discrete transform techniques. *American Journal of Alzheimer's Disease & Other Dementias®*, 31(3):282–293, 2016.

[37] C. Jongkreangkrai, Y. Vichianin, C. Tocharoenchai, H. Arimura, A. D. N. Initiative, et al. Computer-aided classification of Alzheimer's disease based on support vector machine with combination of cerebral image features in MRI. In *Journal of Physics: Conference Series*, volume 694, page 012036. IOP Publishing, 2016.

[38] J. Zhang, Y. Gao, Y. Gao, B. C. Munsell, and D. Shen. Detecting anatomical landmarks for fast Alzheimer's disease diagnosis. *IEEE Transactions on Medical Imaging*, 35(12):2524–2533, 2016.

[39] I. Beheshti, H. Demirel, H. Matsuda, A. D. N. Initiative, et al. Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. *Computers in biology and medicine*, 83:109–119, 2017.

[40] L. Khedher, I. A. Illán, J. M. Górriz, J. Ramírez, A. Brahim, and A. Meyer-Baese. Independent component analysis-support vector machine-based computer-aided diagnosis system for Alzheimer's with visual support. *International journal of neural systems*, 27(03):1650050, 2017.

[41] C. Fang, C. Li, M. Cabrerizo, A. Barreto, J. Andrian, D. Loewenstein, R. Duara, and M. Adjouadi. A novel Gaussian discriminant analysis-based computer aided diagnosis system for screening different stages of Alzheimer's disease. In *Bioinformatics and Bioengineering (BIBE), 2017 IEEE 17th International Conference on*, pages 279–284. IEEE, 2017.

[42] R. K. Lama, J. Gwak, J.-S. Park, and S.-W. Lee. Diagnosis of Alzheimer's disease based on structural MRI images using a regularized extreme learning machine and PCA features. *Journal of healthcare engineering*, 2017, 2017.

[43] M. Liu, J. Zhang, E. Adeli, and D. Shen. Landmark-based deep multi-instance learning for brain disease diagnosis. *Medical image analysis*, 43:157–168, 2018.

[44] F. Li, M. Liu, A. D. N. Initiative, et al. Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. *Computerized Medical Imaging and Graphics*, 2018.

[45] J. Ramírez, J. Górriz, D. Salas-Gonzalez, A. Romero, M. López, I. Álvarez, and M. Gómez-Río. Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features. *Information Sciences*, 237:59–72, 2013.

[46] T. Tong, K. Gray, Q. Gao, L. Chen, D. Rueckert, A. D. N. Initiative, et al. Multi-modal classification of Alzheimer's disease using nonlinear graph fusion. *Pattern recognition*, 63:171–181, 2017.

[47] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng. Early diagnosis of Alzheimer's disease with deep learning. In *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, pages 1015–1018. IEEE, 2014.

[48] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative, et al. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, 101:569–582, 2014.

[49] X. Zhu, H.-I. Suk, L. Wang, S.-W. Lee, D. Shen, A. D. N. Initiative, et al. A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. *Medical image analysis*, 38:205–214, 2017.

[50] T. D. Vu, H.-J. Yang, V. Q. Nguyen, A.-R. Oh, and M.-S. Kim. Multimodal learning using convolution neural network and sparse autoencoder. In *Big Data and Smart Computing (BigComp), 2017 IEEE International Conference on*, pages 309–312. IEEE, 2017.

[51] A. Mikhno, P. M. Nuevo, D. P. Devanand, R. V. Parsey, and A. F. Laine. Multimodal classification of dementia using functional data, anatomical features and 3D invariant shape descriptors. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pages 606–609. IEEE, 2012.

[52] Y. Cui, B. Liu, S. Luo, X. Zhen, M. Fan, T. Liu, W. Zhu, M. Park, T. Jiang, J. S. Jin, et al. Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. *PloS one*, 6(7):e21896, 2011.

[53] O. B. Ahmed, J. Benois-Pineau, M. Allard, G. Catheline, C. B. Amar, A. D. N. Initiative, et al. Recognition of Alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and multiple kernel learning. *Neurocomputing*, 220:98–110, 2017.

[54] E. E. Bron, M. Smits, J. M. Papma, R. M. Steketee, R. Meijboom, M. De Groot, J. C. van Swieten, W. J. Niessen, and S. Klein. Multiparametric computer-aided differential diagnosis of Alzheimer's disease and frontotemporal dementia using structural and advanced MRI. *European radiology*, 27 (8):3372–3382, 2017.

[55] W. Chau and A. R. McIntosh. The Talairach coordinate of a point in the MNI space: how to interpret it. *Neuroimage*, 25(2):408–416, 2005.

[56] D. Collins, P. Neelin, T. Peters, and A. Evans. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of computer assisted tomography*, 18(2):192–205, 1994.

[57] M. Cognition and Brain Sciences Unit. University of Cambridge. The MNI brain and the Talairach atlas, 2009. URL `http://imaging.mrc-cbu.cam.ac.uk/imaging/MniTalairach#Introduction`. Last access: September 28th, 2018.

[58] C. J. Holmes, R. Hoge, L. Collins, R. Woods, A. W. Toga, and A. C. Evans. Enhancement of MR images using registration for signal averaging. *Journal of computer assisted tomography*, 22(2): 324–333, 1998.

[59] J. Zhang, M. Liu, L. An, Y. Gao, and D. Shen. Alzheimer's disease diagnosis using landmark-based features from longitudinal structural MR images. *IEEE journal of biomedical and health informatics*, 21(6):1607, 2017.

[60] P. Morgado, M. Silveira, and J. S. Marques. Diagnosis of Alzheimer's disease using 3D local binary patterns. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 1(1):2–12, 2013.

[61] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International journal of computer vision*, 43(1):29–44, 2001.

[62] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International journal of computer vision*, 62(1-2):61–81, 2005.

[63] C. Schmid. Constructing models for content-based image retrieval. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2001.

[64] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[65] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[66] J. Xu, L. Xiang, R. Hang, and J. Wu. Stacked sparse autoencoder (SSAE) based framework for nuclei patch classification on breast cancer histopathology. In *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, pages 999–1002. IEEE, 2014.

[67] Y. Ju, J. Guo, and S. Liu. A deep learning method combined sparse autoencoder with SVM. In *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2015 International Conference on*, pages 257–260. IEEE, 2015.

[68] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer, 2012.

[69] C. Tao, H. Pan, Y. Li, and Z. Zou. Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification. *IEEE Geoscience and remote sensing letters*, 12(12):2438–2442, 2015.

[70] A. Shmilovici. Support vector machines. In *Data mining and knowledge discovery handbook*, pages 231–247. Springer, 2009.

[71] C. J. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[72] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[73] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[74] S. Boughorbel, J.-P. Tarel, and N. Boujemaa. Generalized histogram intersection kernel for image recognition. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 3, pages III–161. IEEE, 2005.

[75] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29 (2-3):131–163, 1997.

[76] I. Rish et al. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York, 2001.

[77] D. D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.

[78] M. F. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, 6(4):525–533, 1993.

[79] S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91, 2006.

[80] ADNI. Alzheimer's Disease Neuroimaging Initiative, 2017. URL `http://adni.loni.usc.edu`. Accessed: 25/04/2018.

[81] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. J. Hudspeth. *Principles of neural science*. McGraw-Hill Education / Medical, 2012.

[82] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al. A practical guide to support vector classification. 2003.

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