INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Automated Diagnosis of Alzheimer's Disease using PET Images

A study of alternative procedures for feature extraction and selection

Pedro Miguel Maravilha Morgado
(BSc)

Dissertation in support of the candidature for the Master degree in

Electrical and Computer Engineering

Jury

President:	Prof. Carlos Filipe Gomes Bispo
Advisor:	Prof. Maria Margarida Campos da Silveira
Co-Advisor:	Prof. Jorge dos Santos Salvador Marques
Reader:	Prof. Ana Luísa Nobre Fred

September 2012

Success is not final, failure is not fatal: it is the courage to continue that counts.

Winston Churchill

Acknowledgments

This work was funded by Fundação para a Ciência e Tecnologia (FCT/MCTES) through the ADIAR project (PTDC/SAU-ENB/114606/2009).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904).

I would also like to express my gratitude to my advisors, Professor Margarida Silveira and Professor Jorge Salvador Marques, for their valuable and constructive suggestions during the planning and development of this work.

My grateful thanks are also extended to Dr Durval Campos Costa and Eduardo Bicacro for the acquisition of Eye Tracking data, which revealed to be a precious input to this thesis.

Finally, I wish to thank my family and friends for their continuous support and encouragement.

To all of you, I send my sincere acknowledgments.

Abstract

Currently, there is no cure for Alzheimer's disease, but its early detection is essential to an effective treatment, slowing down the progression of symptoms. Consequently, the development of automatic diagnostic tools, which use as principal source of information three-dimensional images of the brain, has attracted great interest in recent years. This work focused on PET images and studied alternatives to two of the main building blocks of a computerized diagnostic system: the extraction and selection of features. Regarding the common approach based on Voxel Intensities, the FDG-PET image was studied for different scales and resolutions. In addition, the use of a measure of local contrast was also tested, as well as the widely known texture descriptor, Local Binary Patterns, to which a novel extension to three dimensional data was proposed. As regards selection, a new method based on data acquired by the *Eye Track* technology during the inspection of PET images by an expert physician was proposed. The aim of this method is to model the behavior of the gaze over time, and use the model to select the features that the expert found most interesting. Moreover, other more conventional methods based on correlation measures and mutual information were also studied. The Support Vector Machine classifier was used to perform binary classifications among AD patients, patients with Mild Cognitive Impairment and a control group (in a dichotomous fashion), obtaining comparable or superior performances to those achieved by most systems found in the literature.

Keywords

Alzheimer's Disease, Computer Aided Diagnosis, Positron Emission Tomography, Feature Extraction, Feature Selection, Eye Tracking

Resumo

Actualmente, não existe cura para a doença de Alzheimer, mas o seu diagnóstico precoce é essencial para um tratamento eficaz, retardando o progresso dos sintomas. Como consequência, o desenvolvimento de sistemas automáticos de diagnóstico que usam como principal fonte de informação imagens tridimensionais do cérebro tem despertado grande interesse nos últimos anos. Este trabalho focou-se em imagens PET e estudou alternativas para dois dos principais blocos constituintes de um sistema computorizado de diagnóstico: a extração e a seleção de *features*. No que respeita a abordagem comum baseada nas intensidade dos voxeis, diferentes resoluções da imagem FDG-PET foram estudadas. Além disso, foi também testado a utilização de uma medida de contraste local e o conhecido descritor de texturas, *Local Binary Patterns*, para o qual foi proposto uma nova extensão para três dimensões. Quanto à seleção de *features*, foi proposto um novo método baseado em dados adquiridos pela tecnologia *Eye Track* durante a inspeção de imagens PET por parte de um especialista. O objetivo deste método é modelar o comportamento do olhar ao longo tempo e usar este modelo para selecionar as *features* que o especialista achou de maior interesse. Estudou-se ainda métodos mais convencionais baseados em medidas de correlação e de informação mútua. O algoritmo *Support Vector Machine* foi utilizado para realizar classificações binárias entre pacientes com Alzheimer, pacientes com Défice Cognitivo Ligeiro e um grupo de controlo (dois a dois), obtendo-se desempenhos comparáveis ou superiores aos alcançados por sistemas semelhantes encontrados na literatura.

Palavras Chave

Doença de Alzheimer, Diagnóstico Assistido por Computador, Tomografia por Emissão de Positrões, Seleção de *Features*, Extração de *Features*, *Eye Tracking*,

Contents

List of Figures

List of Tables

Acronyms

ACC Accuracy

AD Alzheimer's Disease

ADNI Alzheimer's Disease Neuroimaging Initiative

CAD Computer Aided Diagnosis

CDR Clinical Dementia Rating

CN Cognitive Normal

CV Cross Validation

ETDS Eye Track Driven Selection

LBP Local Binary Patterns

LVAR Local Variance

MCI Mild Cognitive Impairment

MI Mutual Information

MIM Mutual Information Maximization

MMSE Mini Mental State Examination

mRMR Minimal Redundancy Maximal Relevance

PBCC Point Biserial Correlation Coefficient

PET Positron Emission Tomography

RBF Radial Basis Function

SENS Sensitivity

SPEC Specificity

SPECT Single-Photon Emission Computed Tomography

SVM Support Vector Machine

TD-ETDS Time Dependent Eye Track Driven Selection

TI-ETDS Time Independent Eye Track Driven Selection

VI Voxel Intensity

1

Introduction

Contents

1.1 Motivation

1.1.1 Alzheimer's Disease and Its Impact on Society

Alzheimer's Disease (AD) is a neurological disorder that mostly affects people over 65 years old and whose incidence rate grows exponentially with age, almost doubling in every 5 years [1]. Although it has been described for the first time more than 100 years ago, by Alois Alzheimer, only in the last 30 years its causes, symptoms, risk factors and treatment have been intensively investigated. Still, apart from a few exceptions, the factors that trigger the onset of AD remain unknown [2]. It is a progressive disease, meaning that it worsens over time, and for which there is currently no cure, leading eventually to death. The very early stages are often mistakenly confused with the normal process of ageing or linked to stress and it is often characterized by episodic losses of short term memory and difficulty to grasp new ideas. This preclinical stage is also known as Mild Cognitive Impairment (MCI). As the brain damage progresses, other cognitive impairments appear and the disease becomes obvious. In the late stages, individuals are completely dependent on caregivers even for the most basic daily tasks such as eating, bathing or dressing. Moreover, motor skills are affected and patients become more vulnerable to infections. Pneumonia, a lung infection, is one of the most frequent direct causes of death [2–4].

In 2010, nearly 35.6 million people worldwide were living with Alzheimer but this figure is rapidly increasing. In fact, the prevalence of AD is estimated to almost double in every 20 years, reaching a total of 65.7 million in 2030 and 115.4 million in 2050 [3]. This growth will be more significant in low and middle income countries [3] as shown in Figure 1.1, and it is driven mainly by the demographic ageing and population growth. For instance, in regions such as Central Latin America, North America and Middle East, the expected growth between 2010 and 2050 is over 400% [3]. Also, the number of deaths related to Alzheimer's is still experiencing a marked increase, contrarily to other major causes whose numbers are declining. This fact is not surprising since AD remains incurable and the number of people affected is not showing signs of slowing down. As a consequence, AD is already one of the most important causes of death, particularly in developed countries, ranking fifth in the United States

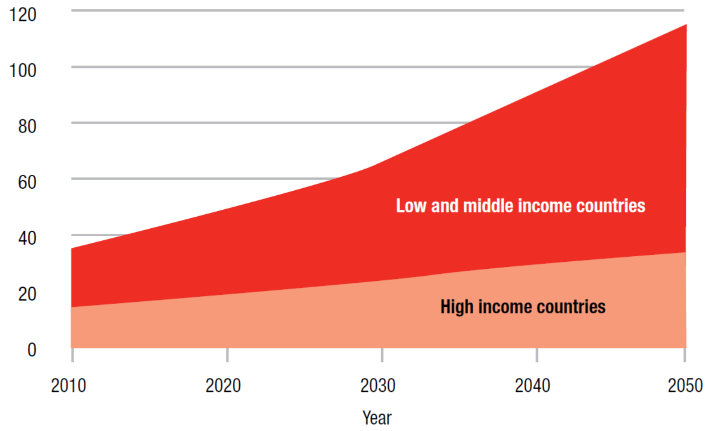


Figure 1.1: Estimated number of people with dementia (in millions) until 2050 in high income, and low and middle income countries. Source: World Alzheimer Report 2009 [3].

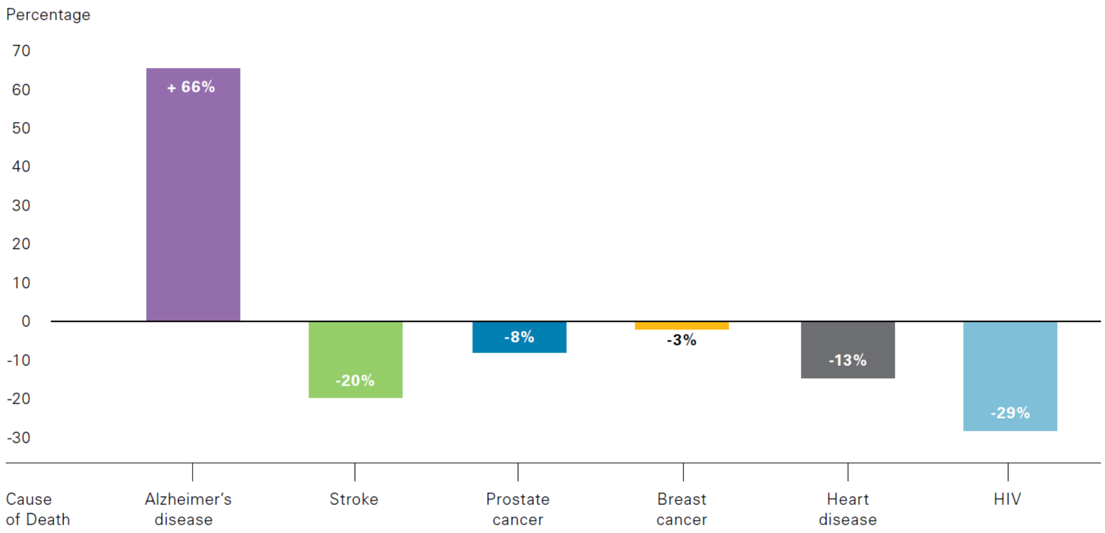
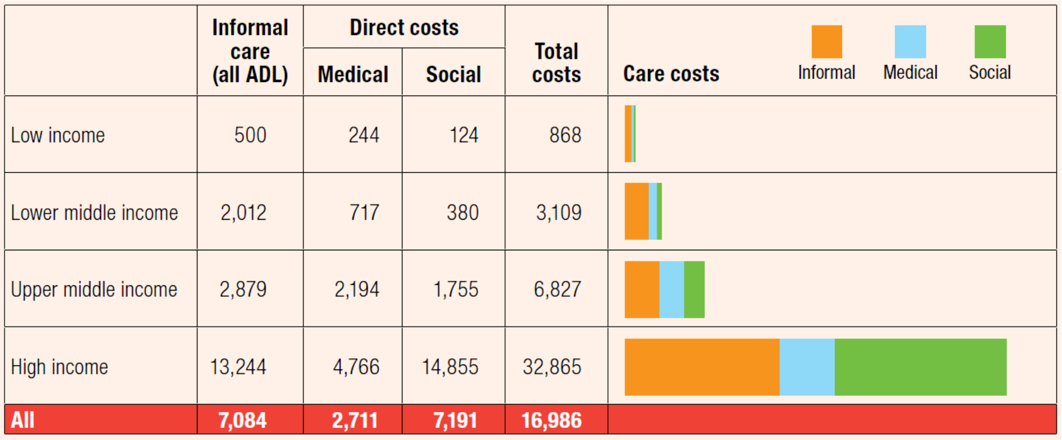


Figure 1.2: Percentage changes in selected causes of death between 2000 and 2008 in the United States. Source: Alzheimer's Association Facts and Figures 2012 [2]

for those aged 65 or older [2]. Figure 1.2 compares changes in mortality between the years of 2000 and 2008 for several diseases.

Beyond the obvious impact on the actual patients, the burden of Alzheimer's disease also falls on close relatives whose own health and life quality become seriously compromised as well [5]. As mentioned before, as the damage spreads throughout the brain, the patients lose their ability to perform basic daily tasks, therefore becoming completely dependent on caregivers. Also, due to some characteristic symptoms of AD, such as memory loss, irritability and personality changes, caring exposes the caregiver to high levels of stress and it is an expensive and time-consuming task [2]. Alzheimer's Disease Association estimates that 17.4 billion hours of unpaid care were provided and more than US$200 billion were spent for caring purposes only in the United States in 2011 [2]. A different study estimated that in 2010 the worldwide costs associated with AD were over US$600 billion, a figure higher than the total annual revenues of the 2010 world's largest company, Wal-Mart, and higher than the gross domestic product of the 18th largest economy, Indonesia, and even more

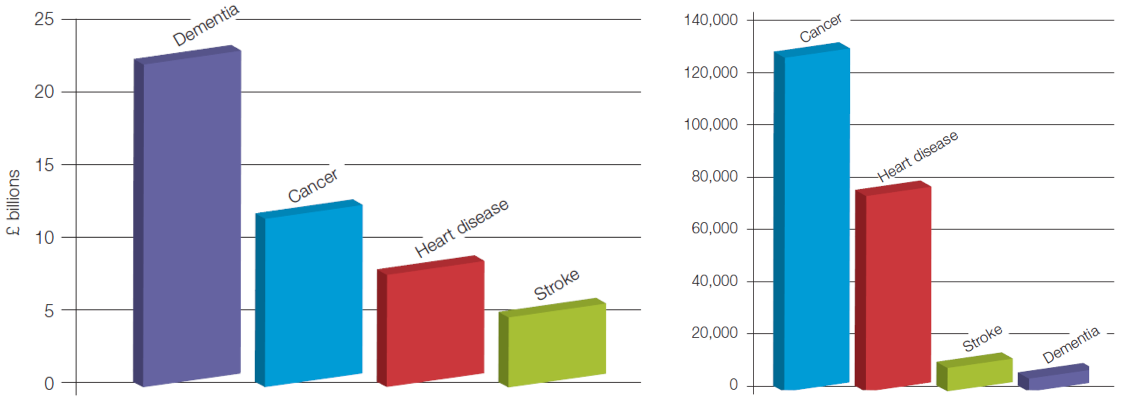


	Informal care (all ADL)	Direct costs		Total costs	Care costs 🟧 Informal 🟦 Medical 🟩 Social
		Medical	Social		
Low income	500	244	124	868	
Lower middle income	2,012	717	380	3,109	
Upper middle income	2,879	2,194	1,755	6,827	
High income	13,244	4,766	14,855	32,865	
All	**7,084**	**2,711**	**7,191**	**16,986**	

Figure 1.3: Costs per person in different World Bank income groups (US$). Source: World Alzheimer Report 2010 [4].

alarming, that these numbers are doomed to increase due to the increasing number of sufferers [4]. Figure 1.3 presents the worldwide economic impact of dementia, showing the costs per patient in different types of countries, from high income to low income, and highlighting cost sources such as the value of informal unpaid caring, and medical and social costs. Figure 1.3 clearly states that high income countries spend more money caring for their patients, as expected.

A comparison between the economic impact of dementia and other common diseases was conducted in the UK by the Health Economics Research Centre from Oxford University [6], revealing that health and social costs of dementia are twice as much as the ones of cancer and three times as much as the ones of heart disease (Figure 1.4(a)). Even so, the funds allocated for research of dementia is 26 times smaller than the ones for cancer investigation and 15 times smaller when compared to heart disease (Figure 1.4(b)).



(a) Annual Costs in UK

(b) Investment (£) in research per £1 million spent in social and medical costs

Figure 1.4: Comparison between health and social care costs with research funding by disease. Source: Dementia 2010 [4].

All previously mentioned studies contribute to help us capturing the real social and economic burden that AD poses to present and future generations. Thus, it is of greatest interest to intensify the research on prevention, diagnosis, treatment and care of AD in order to prevent the exponential growth both of the number of deaths and of the economic impact attached to Alzheimer's disease.

1.1.2 The role of PET

The most common criteria used for the diagnosis of Alzheimer's disease were published in 1984 by the Alzheimer's Association and the National Institute of Neurological Disorders and Stroke, but recently, they have been updated to include the latest scientific advances [2]. The diagnosis is often performed by the primary care physician and is based on the cognitive and behavioral history of the patient, which is assessed based on direct interviews with the patient himself and relatives, and through the usage of several cognitive, physical and neurological tests such as the Mini Mental State Examination (MMSE) [2]. The severity of the symptoms of dementia is often measured in a scale known as the Clinical Dementia Rating (CDR). Neuroimaging techniques are also used, when available, to increase the confidence of diagnosis, since a definite diagnosis is only possible post-mortem in histological examination.

Several neuroimaging techniques, such as Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI) and Single-Photon Emission Computed Tomography (SPECT), have shown to hold precious information about the presence of AD, even in its early stages, so precious that it is possible to build a fully automated diagnostic system using only the three-dimensional brain images and still achieve good performance results. Chapter 2 will present several examples of such systems for the Computer Aided Diagnosis (CAD) of AD. In the current work, the FDG-PET neuroimaging technique was explored.

PET is a nuclear medicine imaging technique whose operating principle relies on the detection of pairs of gamma rays emitted by a positron-emitting radionuclide, also known as tracer, which is introduced into the body on a biologically active molecule. When FDG, which is an analogue of glucose, is used as the biologically active molecule, the scan produces an image that measures the regional glucose uptake, and thus when a tomography is performed on the brain, the subsequent image measures the brain metabolism directly, allowing for the detection of what is believed to be the earliest observable anomaly associated with AD: the reduction of the metabolism in certain areas of the brain [8]. In fact, in the last 20 years, research on the diagnostic value of FDG-PET for AD consistently showed a reduction of the cerebral metabolic rate for glucose (CMR$_{glc}$) and perfusion present in several structures of the brain, such as the posterior cingulate and temporoparietal association cortices, largely sparing the basal ganglia, thalamus, cerebellum and cortex mediating primary sensory and motor functions [9,10]. In addition, quantitative studies have found significant absolute reductions in CMR$_{glc}$, achieving values between 7% and 17% in the medial temporal lobe, 8% in the lateral temporal and 23% in the posterior cingulate cortices [8]. An example of typical FDG-PET findings related with AD is illustrated in Figure 1.5.

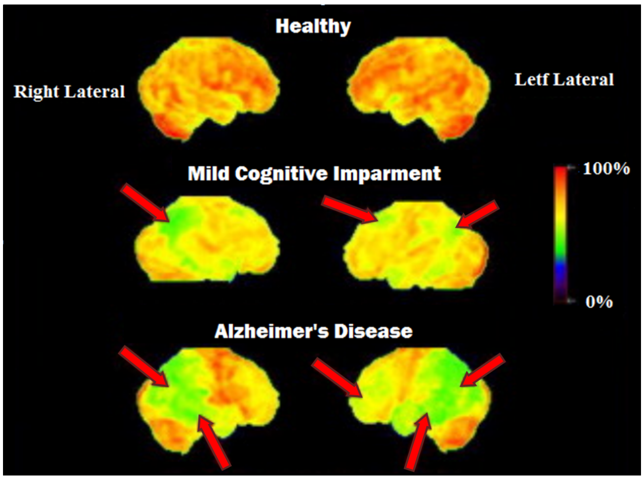


Figure 1.5: Typical findings in healthy subjects, patients with MCI and patients with AD. Surface projections of cerebral glucose metabolism in both hemispheres (Red/yellow – Normally high cerebral glucose metabolism; Green – Abnormally low metabolism). Source: Drzezga [7].

1.2 Proposed Approach

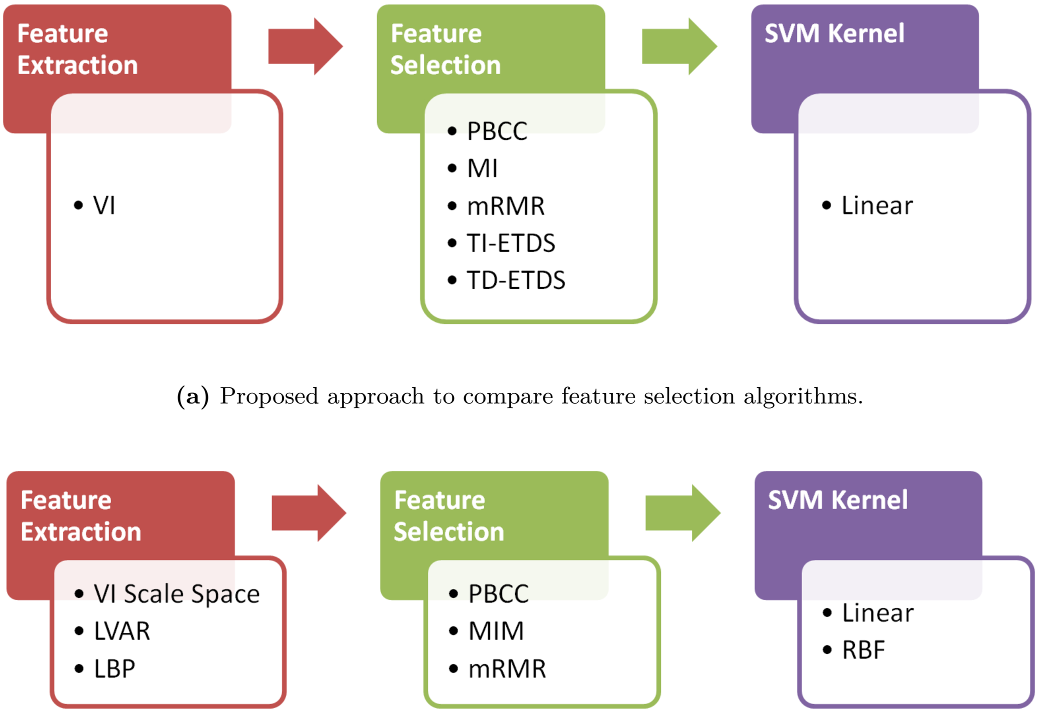
Neuroimaging data used in this thesis were retrieved from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [11], a large study whose objective has been to gather different types of predictive biomarkers, not only from subjects suffering from AD but also from MCI patients and normal controls (Cognitive Normal (CN)). FDG-PET is among the collected biomarkers. In addition, all images in the ADNI database were acquired under standardized conditions and had also been subject to a preprocessing step to ensure that all scans are represented in a common space.

Predicting mild cognitive impairment and its evolution to Alzheimer's has sparked a growing interest in this research field (CAD of AD). Therefore, not only the classification of Alzheimer's disease against normal controls was studied, but also the automated diagnosis involving patients suffering from MCI was carried out, adding to a total of three different classification tasks: AD vs. CN, MCI vs. CN and AD vs. MCI.

Various challenges arise in most real world classification problems, which increase substantially their inherent difficulty. In this specific context, one of the most significant adversities is the high dimensionality of the input brain image, which is the central source of information for the classification task. Actually, it is known that the performance of classification algorithms is often jeopardized by high dimensional input patterns, a phenomenon known as the *curse of dimensionality.* As a consequence, a preprocessing stage is often exploited to reduce the number of features that will describe each subject in the learning phase.

In this thesis, in addition to ranking methods based on the Point Biserial Correlation Coefficient (PBCC) and Mutual Information (MI) between each feature and the class, which were already extensively studied for the CAD of AD, three different methods were tested. One is also based on MI but accounts for the redundancy between features and it is known as Minimal Redundancy Maximal Relevance (mRMR). It is an established algorithm for feature selection but was never used in this context. The other two are based on a model built from a physician's eye tracking procedure. The difference between them is that one ignores the time sequence through which the physician visited each region of the brain, while the other does not. They were called Time Independent Eye Track Driven Selection (TI-ETDS) and Time Dependent Eye Track Driven Selection (TD-ETDS), respectively. In order to compare all feature selection algorithms, they were all tested with Voxel Intensity (VI) features, and the classification was carried out by a Support Vector Machine (SVM) with linear kernel. Parameter optimization was performed within a nested Cross Validation (CV) procedure. This first experiment is summarized in Figure 1.6(a).

Alternative feature extraction techniques, beyond Voxel Intensity, were also assessed, namely Local Binary Patterns (LBP) and Local Variance (LVAR). LBPs are commonly used in image processing for texture analysis and, up to the present date, were only applied once in the classification of dementia, but not specifically of AD. LVAR, on the other hand, captures the local contrast within a given image. Regarding the starting VI features, PET images were also studied at different scales and resolutions using a Gaussian pyramid representation of the image's scale-space. For each type of feature (LBPs, LVAR and VI) the best combination of feature selection algorithm and SVM kernel was sought, in order to achieve the best possible generalization accuracy. However, selection algorithms were restricted to a pool containing PBCC, MI and mRMR for reasons that will become clear in chapter 6. Also, both linear and Radial Basis Function (RBF) were tested as kernel functions for the SVM algorithm, the algorithm utilized to learn all three classification problems. Once again, nested cross-validation was used to tune the existing parameters. The experiment just described is schematized in Figure 1.6(b).



(a) Proposed approach to compare feature selection algorithms.

(b) Proposed approach to compare feature extraction algorithms.

Figure 1.6: Proposed approach. Each implemented classifier is composed by one of the listed components in each building block: feature extraction, feature selection and SVM kernel.

1.3 Original Contributions

The present work brings innovative methods to the context of automated diagnosis of Alzheimer's disease, specifically, to the feature extraction and to the feature selection phases. In the feature extraction stage, most research has been focused directly on Voxel Intensity features [12–20] and even when other types of features were used, such as deformation fields [21], shape descriptors [22] or LBPs [23], the majority was restricted to specific regions of interest. Herein, two different types of features, LVAR and LBPs, were studied and implemented so as to ensure full coverage of the brain. In addition, the feature LVAR had never been used for the CAD of AD, contrary to LBPs which were recently used (in the present year) for the diagnosis of several types of dementia [23]. Nevertheless, that work did not focus specifically on AD or PET images, and exploited the original two-dimensional version. In this thesis, a novel extension of LBPs to three dimensions is proposed, which differs from known 3D LBP schemes because no approximation to the original concepts was made. The most important contribution of the present work to the 3D extension is the redefinition of the uniformity concept which allows for its generalization to any dimension, while achieving the same results in 2D. In addition, a new approach to three-dimensional rotation invariance is also proposed.

Regarding feature selection algorithms, the already established procedure mRMR [24] was brought for the first time to this research field. Also, a novel scheme based on data collected from a physician's eye tracking procedure was proposed. In a previous work, Bicacro et al. [18] carried the aforementioned procedure and implemented a feature selection algorithm based on the resulting data, a series of points in the brain fixated by the physician while performing the diagnosis. However, they ignored the path through which the physician analyzed the brain. Here, this information was not discarded bringing forth a new feature selection technique.

1.4 Thesis Outline

The remainder of this dissertation is organized in the following way. First, the most important contributions to the computerized diagnosis of AD will be highlighted in chapter 2. Then, in chapter 3, each type of feature used in this study will be carefully exposed, giving special attention to the concepts underlying their implementation. Afterwards, the feature selection algorithms exploited in the dimensionality reduction stage will be presented in chapter 4, describing not only the concepts but also how they relate to each other and the reasons that motivate their utilization. Next, in chapter 5, the fundamentals behind the SVM algorithm and the nested CV procedure will be covered. Chapter 6 will be focused on experimental results, starting with the description of the image database used in the present work, and then presenting the experimental design and the actual results. Finally, chapter 7 concludes the dissertation.

2

State of the Art

Contents

2.1 Introduction

In the last decade, there has been a growing interest into developing an automatic CAD system that reliably differentiate subjects that suffer from medical conditions such as AD and MCI, from healthy controls or even from subjects with different kinds of cognitive dementia that could be confused with AD, like Frontotemporal Lobar Degeneration (FTLD). An important motivation relates to the fact that today's diagnostic procedures are highly dependent on the physician's radiological expertise (when examination of brain images is required) and are very time-consuming, taking typically a few weeks to complete the evaluation [25]. Also, the early diagnosis of AD, which is essential to improve the efficacy of current treatments, is very complex because no characteristic pattern of brain degeneration is well-defined, and therefore automated tools may allow a more sensitive analysis and improve diagnostic accuracy.

In the remaining of this chapter, a review of the main contributions and the principal trends in this research field will be highlighted. First, in section 2.2, an overview of the most frequently used biomarkers will be given. Then, in section 2.3, the types of features that were used in the literature to successfully distinguish AD patients, MCI patients and normal controls will be described. Next, a summary of feature selection techniques utilized for the CAD of AD will be presented in section 2.4, and a summary of learning and performance evaluation algorithms in section 2.5. In the last two sections, 2.6 and 2.7, relevant works will be shortly summarized, focusing on the innovative ideas that they introduced.

2.2 Biomarker

Neuroimaging data hold valuable information for the diagnosis of AD and related disorders, as discussed in the previous chapter, and, consequently, it has been seen as the central biomarker to the CAD system. Although this study focuses solely on PET images, other neuroimaging techniques, namely SPECT and MRI, were explored in the literature (see Tables 2.1 and 2.2 for examples), either as the unique source of information [12,14,21], or combined with each other [20,22] or even combined with other clinically relevant data, such as APOE genotype information [16,26] or the result of MMSE test [27,28].

The integration of several neuroimaging techniques is rational because they trace different biomarkers and, thus, complementary information can be retrieved. For instance, FDG-PET measures glucose uptake, PIB-PET measures brain amyloid levels, SPECT measures cerebral blood flow, and MRI is a structural image of the brain. Features associated with Cerebrospinal Fluid (CSF) were also found in the literature [20]. CSF influences the automatic regulation of cerebral blood flow and, therefore, its assessment provides useful information for the diagnosis of AD.

2.3 Features and Feature Transformations

Features retrieved from the brain image play an important role in the success of a given system and a considerable effort has been made to find more discriminant features. In what concerns the type of feature, previous studies can be cataloged in two distinct classes: those who use Regions of Interest (ROI) [19, 22, 29] and those who use the whole brain [17, 18, 30].

In defense of the first class, several arguments related to previous knowledge about the disease can be presented, since several studies had previously identified the regions of the brain that are mostly affected by the disease. Moreover, using only a fraction of the brain image reduces significantly the dimensionality of the feature vectors, therefore alleviating the *curse of dimensionality* (see section 4.1 for a more comprehensive description of this phenomenon). In addition, highly specific characteristics of those ROIs, such as the volume of gray matter tissue [22,31] or the shape of the hippocampus [22,32], are often used as features reducing even more the input dimensionality.

However, this approach has its own disadvantages. It requires the choice, in advance, of the ROIs to be studied, and the manual or semi-automatic extraction of those regions is unavoidable, which is a difficult, time-consuming and user dependent task. This is the reason why CAD systems that build their classifier over the whole brain, without further knowledge, also share the limelight of recent research. In this second category, VI features are the most common, regardless of their meaning that depends on the neuroimaging modality. Nevertheless, features obtained from transformations of the brain volumes, such as Histograms of Gradient Magnitude and Orientation [33], 3D Haar-like features [33], deformation fields [21] or Normalized Mean Square Error [30], have been reported in previous studies in order to capture complementary information.

2.4 Feature Selection and Dimensionality Reduction

Dimensionality reduction is one additional component common to most CAD systems both for the ones that use the whole brain and for those that use ROIs. The grounds for this step are linked, once again, to the high dimensionality, low sample size problem. To get a rough idea of the gap between the number of features and the sample size, in the whole brain based systems, the number of voxels easily exceeded tens or even hundreds of thousands, and in the ROIs based systems, this number, despite being smaller, reached a few hundreds in the simplest setting found in the literature. On the other hand, the cardinality of datasets available for study was usually smaller than 200.

Distinct approaches have been tested regarding this problem, including methods that study linear combinations of the original variables like Principal Component Analysis (PCA) [14, 16], Linear Discriminant Analysis (LDA) [15] or Nonnegative Matrix Factorization (NMF) [34], and feature selection procedures, more specifically ranking algorithms that assign one measure of relevance to each feature to select the most important ones. From the measures of relevance found in the literature, one can highlight the mutual information [18, 33], correlation coefficients [18, 30, 33], the Fisher Discriminant Ratio (FDR) [34] and the absolute value of the two-sample t-test statistic [30].

The main advantage of the first type of methods (PCA and LDA) is that they are able to account for combinations of the input features during the process of dimensionality reduction, while ranking methods only look at one feature at a time. Note that there are selection procedures which do not belong to the class of ranking algorithms and that account for relations between features, such as mRMR, but none was previously used for the CAD of AD. The difference between PCA and LDA is that the latter tries to find a lower subspace where instances belonging to different classes are separated, while the former does not take the output label into account, which makes LDA more suitable for the problem at hand. The main disadvantage of these techniques is the higher computational needs when compared to ranking algorithms.

2.5 Learning Machine and Performance Evaluation

The final component that all CAD systems share is the learning machine. Generally, supervised learning machines can be grouped into two classes: a generative approach that tries to learn the probability functions behind the problem and then classifies a given pattern according to the most probable output label, and a discriminative approach that focus directly on the prediction. The small sample size problem makes the first approach, based on generative models, to become unreliable because the estimation of the parameters would not be trustworthy. This is the reason why most studies used the second approach, i.e., used discriminative models. The most frequently used learning algorithm was SVM mainly due to its great robustness to problems in high dimensional spaces. SVM was also exploited in the current work and it will be further detailed in section 5.2.

Still, experiments were conducted with different classifiers, such as Adabost [17] which is a Boosting algorithm that performs classification based on a combination of multiple simple classifiers, called "weak" classifiers, Random Forests [31] which is also an ensemble learning machine where each classifier is a binary tree, and even with the Naive Bayes [15] and Maximum Likelihood [23] classifiers which are based on generative models. The last two classifiers relied heavily on the dimensionality reduction stage, so that the training sample size would become greater than the number of parameters to estimate.

Regarding performance evaluation, cross-validation was the most frequently used algorithm [17,30, 34] due to the low number of subjects present in most datasets. It was also used in the present study and, therefore, it will be discussed in section 5.3. Apart from cross-validation, other techniques have been reported. In a study where the cardinality of the dataset was higher than normal (around 200 for each class), the starting data were divided in a training set for learning and a test set for performance evaluation [26]. On the contrary, in studies where the cardinality of the dataset was even lower than normal (less than 30 for each class), some authors opted to use a bootstrapping technique [22], which is known to achieve more stable results than cross-validation at the cost of attaining optimistically biased estimates. In the next section, where different works will be addressed one by one, nothing will be said about evaluation techniques, unless an algorithm different from cross-validation is used.

2.6 Important Contributions

Henceforth, a chronological review of the most important contributions to the automatic diagnosis of AD will be presented. Tables 2.1 and 2.2 summarize the content of the section that now starts.

Before the work of Stoeckel et al. in 2001 [12], most research was not focused on classifying single subjects, but instead on finding regions of the brain where different blood flow patterns could distinguish AD from normal controls. In this work, Stoeckel et al. formulated the problem as a supervised learning one, in order to be able to do future predictions of single individuals using SPECT images. The voxel intensities of the brain volume were used directly. In order to circumvent the small sample size problem, they reduced the number of features using lower resolution equivalents of the original image, which were computed by smoothing with a mean filter, followed by a sub-sampling step. They tested the Nearest Mean Classifier (NMC) and the Pseudo Fisher Linear Discriminant (PFLD) as learning machines. The best result was achieved using PFLD and a sub-sampling factor of 4, yielding 89.9% of accuracy, 82.8% of sensitivity and 94.0% of specificity (from now on and to shorten the exposition, these three results will be given in the following format "89.9% (82.8, 94.0)%").

Stoeckel's follow-up work, in collaboration with Glenn Fung, in 2005 [13] introduced a novel method to perform simultaneously feature selection and classification, based on the SVM algorithm and regularization theory. It was named Contiguous Linear SVM. This algorithm seeks to select the most relevant "areas" of the brain, instead of just the most important voxels as a simple SVM based on the l_1-norm would have done. The results of the new CAD system, 86.0% (84.4, 90.9)%, although lower than the estimates of the previous study, were obtained in a different dataset. Actually, the first CAD system, using the PFLD classifier, was evaluated in the new dataset yielding worse results, 83.3% (82.0, 87.5)%.

In 2008, the field of the CAD of AD experienced a clear expansion with far more papers being published from that year forward. Duchesne et al. [21] used an SVM to classify MRI images, but instead of using just the voxel intensities, this group also used deformation fields to include local shape information. They restricted themselves to a large volume of interest (VOI), smaller than the whole brain, centered on the median temporal lobe and used PCA for dimensionality reduction purposes. The CAD system was able to correctly differentiate CN from AD patients with an accuracy of 92.0%. They also tested two other classifiers based on LDA and Quadratic Discriminant Analysis (QDA) but with worse results. In a different work, Góriz et al. [29] proposed a component based SVM algorithm using SPECT images, i.e. the original volumes, after a sub-sampling step, were divided into several regions and for each region a different SVM classifier was learned. The final classification was computed through the sum of votes, each vote weighted by the accuracy obtained in the training set by the corresponding SVM. Still in 2008, Vemuri et al. [26] published their work where, in addition to MRI voxel intensity features, also genotype data (APOE) and demographic features (age and gender) of each subject were used to train a support vector machine, attaining a classification accuracy of 89%. One feature selection step was conducted by training a preliminary SVM and retaining only the most discriminative features. Xia et al. [14] introduced a novel approach to the dimensionality reduction stage. They employed PCA together with a genetic algorithm to find the most discriminative features, and then used the SVM algorithm to tell AD from CN subjects apart. Voxel intensities from FDG-PET images were directly used and the CAD system achieved an accuracy of 90%.

In the following year, Chavez et al. and López et al. published two papers, [30] and [15], respectively. Actually, they were part of the same research group together with Górriz and Illán whose works are also summarized in this section. In the first paper, Chavez et al. used, once again, an SVM and SPECT images to create an automated diagnostic system. They brought to light a new type of feature, named

Normalized Mean Squared Error (NMSE), that captures the similarity of the Regional Cerebral Blood Flow (rCBF) of each subject with the mean rCBF of the normal controls present in the training set. To select just the best features, the two-sample t-test statistic was computed for each feature, and then its absolute value was used to evaluate each feature's utility. They were able to achieve 98% accuracy, but they used the .632+ bootstrap method to estimate it. In the second paper, López et al. utilized multivariate techniques, namely PCA and LDA, to extract a small number of features. They were able to reduce the initial image to less than 10 features and still achieve a good separation ability. Since the number of features was so small, they successfully applied a Bayesian framework for automatic classification. The same method was carried out using two different biomarkers, SPECT and PET images, attaining the best results of 93.41% (94.00, 92.68)% and 98.33% (97.62, 100)%, respectively.

In 2010, Silveira and Marques [17] proposed a boosting algorithm, Adabost, with the goal of selecting the most important features within the training stage. Adabost is a nonlinear, iterative method that associates one "weak" classifier with each feature, a voxel of the original FDG-PET image, and then, in each iteration, one feature is chosen to be part of the final classifier, giving more attention to patterns misclassified by features already selected. This method was used to differentiate CN, MCI and AD subjects and achieved 90.97% accuracy in AD vs. CN, 79.63% in MCI vs. CN and 70.00% in MCI vs. AD.

In the year of 2011, Bicacro et al. [18] introduced an innovative approach to the feature selection stage. The authors implemented a medically informed feature selection technique that they called Eye Track Driven Selection (ETDS). The movement of a physician's eye was recorded, while performing the diagnosis of several patients, and then, using a probabilistic interpretation of the data, the voxels that captured most of the physician's attention were selected randomly. The underlying experiment will be presented in more detail in section 4.4, since an approach built on the same data will be developed in the current thesis. Making use of FDG-PET images and the SVM algorithm, Bicacro et al. achieved 91.4% (90.0, 92.8)% in the AD vs. CN task. They also used the same system to diagnose MCI patients but other feature selection methods, based on correlation coefficients or on mutual information, outperformed ETDS. In a different paper [33], the same research group explored alternative feature extraction methods, namely Histograms of Gradient Magnitude and Orientation (HGMO) and 3D Haar-like features. HGMO capture textural information and Haar features are often used in image processing mainly for object detection purposes. They were able to attain 90.8% accuracy in AD vs. CN, 73.6% in MCI vs. CN and 76.1% in AD vs. MCI task for the 3D Haar-like features and 91.6% in AD vs. CN, 65.5% in MCI vs. CN and 73.6% in AD vs. MCI for HGMO. Gray et al. [19] presented another automated diagnostic tool to discriminate subjects from the AD, MCI and CN groups based on their FDG-PET scans. The authors segmented the original image into 83 anatomical regions (ROIs) and used the average intensity within each region as features. No feature selection algorithm was applied. They obtained comparatively worse results, 81.6% accuracy in AD vs. CN, 70.2% in MCI vs. CN and 68.2% in MCI vs. AD. During the same year, an innovative work carried by Zhang et al. [20] studied the combination of multiple biomarkers, namely MRI, FDG-PET and CSF. As mentioned before, the reasoning behind this study is that potentially non-redundant information can be retrieved from several biomarkers. However, the increased number of features adds to the small sample size problem. The authors circumvented this problem by manually labeling 93 ROIs and then, once again, taking the average intensities within each region as features, both for PET and MRI images. Three features from the CSF biomarker were added to this pool. Also, a t-test based feature selection algorithm was employed and an SVM was exploited for classification purposes yielding 81.6% accuracy on the identification of AD patients and 76.4% on MCI patients against normal controls.

More recently, in 2012, Mikhno et al. [22] also used a combination of features retrieved from different neuroimaging modalities. From MRI images, both Hippocampal volume and shape were taken into account and from the FDG-PET and PIB-PET images, the average intensities within 7 ROIs were computed. The hippocampal shape was captured by the 3D Zernike descriptor which relies on spherical harmonics. Then, to select just the most discriminant features, they used an iterative method based on logistic regression where, in each iteration, the feature that, jointly with the previously selected ones, achieved the highest Area Under the Curve (AUC) was chosen. A bootstrapping technique was applied to estimate the generalization ability of an SVM on three problems, AD vs. CN (98.8% acc.), MCI vs. CN (84.3% acc.) and AD vs. MCI (93.3% acc.). In a different work presented by Oppedal et al. [23], LBPs were used to describe textural information in regions with white matter lesions, which were automatically segmented from MRI images, in order to distinguish patients with dementia from normal controls. This study was not confined to patients suffering from AD and included also patients with Lewy Body Dementia. However, it is worth mentioning since LBPs play an important role in this thesis (see section 3.5 for a comprehensive description of this type of feature). A maximum likelihood framework, assuming normal distributions, was used for classification and the CAD performance was evaluated using the AUC of the Receiver Operating Characteristic (ROC) curve. In the same year, Padilla et al. [34] submitted their work, where the proposed CAD system was applied to two biomarkers independently, SPECT and PET images. First, the Fisher Discriminant Ratio (FDR) was applied to select only the most discriminant voxels. Then, the resultant features were projected onto a low dimensional subspace (a maximum 8 dimensions) using a decomposition technique called NMF. At last, a modified SVM with bounds of confidence was utilized as the classifier, which was able to increase its success rate at the expense of having unclassified patients. This CAD system obtained a performance of 88.6% (87.5, 85.4)% when using PET images and 91.4% (90.6, 92.3)% when using SPECT.

2.7 Summary

The following tables list, chronologically, most studies presented in the current chapter. Table 2.1 summarizes the CAD systems proposed until 2010 and Table 2.2 summarizes the ones proposed after 2011, inclusive.

Table 2.1: Performance of different CAD systems proposed until 2010. (*) – Studies using ADNI data. ($^+$) – Performance evaluation using a bootstraping algorithm. Results for different classification tasks are labeled according to the following codes: 1 – AD vs. CN; 2 – MCI vs. CN; 3 – AD vs. MCI. Acronyms: Accuracy (ACC); Sensitivity (SENS); Specificity (SPEC).

Author(s)	Biomarker(s)	Features	Feature Selection	Learning Algorithm	Participants	Results (%)		
						ACC	SENS	SPEC
Stoeckel et al., 2001 [12]	SPECT	Voxel Intensity	Subsampling	PFLD	29 AD 50 CN	89.9	82.8	94.0
Stoeckel & Fung, 2005 [13]	SPECT	Voxel Intensity	—	Contiguous SVM	99 AD 31 CN	86.0	84.4	90.9
Duchesne et al., 2008 [21]	MRI	Voxel Intensity Deformation Fields	VOI PCA	SVM	75 AD 75 CN	92.0	-	-
Górriz et al., 2008 [29]	SPECT	ROIs: VI	Subsampling	SVM Ensemble	39 AD 41 CN	88.6	-	-
Vemuri et al., 2008 [26]	MRI APOE	Metadata Voxel Intensity APOE	SVM based Wrapper	SVM	190 AD 190 CN	89	86	92
Xia et al., 2008 [14]	FDG-PET	Voxel Intensity	PCA Genetic Algorithm	SVM	80 AD 70 CN	90.0	-	-
Chavez et al., 2009 [30]	SPECT	NMSE	t-test statistic	SVM	41 AD 38 CN	98^+		
López et al., 2009 [15]	SPECT	Voxel Intensity	PCA+LDA	Gaussian Naive Bayes	50 AD 41 CN	93.4	94.0	92.7
López et al., 2009 [15]	PET	Voxel Intensity	PCA+LDA	Gaussian Naive Bayes	42 AD 18 CN	98.3	97.6	100
Illán et al.*, 2010 [16]	FDG-PET APOE	Voxel Intensity	PCA	SVM	95 AD 209 MCI 97 CN	88.2^1 70.2^2	87.8^1 83.5^2	88.6^1 40.9^2
Silveira & Marques*, 2010 [17]	FDG-PET	Voxel Intensity	—	Adabost	74 AD 113 MCI 81 CN	91.0^1 79.6^2 70.0^3	- - -	- - -

Table 2.2: Performance of different CAD systems proposed since 2011. (*) – Studies using ADNI data. ($^{+}$) – Performance evaluation using a bootstraping algorithm. Results for different classification tasks are labeled according to the following codes: 1 – AD vs. CN; 2 – MCI vs. CN; 3 – AD vs. MCI. Acronyms: Accuracy (ACC); Sensitivity (SENS); Specificity (SPEC).

Author(s)	Biomarker(s)	Features	Feature Selection	Learning Algorithm	Participants	Results (%)		
						ACC	SENS	SPEC
Bicacro et al.*, 2011 [18,35]	FDG-PET	Voxel Intensity	ETDS	SVM	59 AD 59 MCI 59 CN	91.4^{1} 68.7^{2} 70.5^{3}	90.0^{1} 66.9^{2} 60.9^{3}	92.8^{1} 70.5^{2} 80.1^{3}
Bicacro et al.*, 2011 [33,35]	FDG-PET	HGMO	Correlation Coefficients Mutual Information	SVM	59 AD 59 MCI 59 CN	91.6^{1} 65.5^{2} 73.6^{3}	89.8^{1} 65.4^{2} 64.4^{3}	92.9^{1} 65.7^{2} 78.8^{3}
Bicacro et al.*, 2011 [33,35]	FDG-PET	3D Haar-like	Correlation Coefficients Mutual Information	SVM	59 AD 59 MCI 59 CN	90.8^{1} 73.6^{2} 76.1^{3}	91.5^{1} 76.0^{2} 62.7^{3}	90.0^{1} 70.0^{2} 83.7^{3}
Gray et al.*, 2011 [19]	FDG-PET	ROIs: Average VI	—	SVM	71 AD 147 MCI 69 CN	81.6^{1} 70.2^{2} 68.2^{3}	82.7^{1} 73.8^{2} 58.3^{3}	80.4^{1} 62.3^{2} 73.0^{3}
Zhang et al.*, 2011 [20]	MRI FDG-PET CSF	ROIs: PET VI ROIs: MRI VI CSF	t-test statistic	SVM	51 AD 147 MCI 69 CN	93.2^{1} 76.4^{2}	93.0^{1} 81.8^{2}	93.3^{1} 66.0^{2}
Mikhno et al., 2012 [22]	MRI FDG-PET PIB-PET	Hippoc. Shape Hippoc. Volume ROIs: Average VI	Logistic Regression AUC	SVM	17 AD 22 MCI 17 CN	98.8^{1+} 84.3^{2+} 93.3^{3+}	99.5^{1+} 82.9^{2+} 93.6^{3+}	98.1^{1+} 85.9^{2+} 93.3^{3+}
Oppedal et al., 2012 [23]	MRI	2D LBPs	—	Maximum Likelihood	52 AD 32 CN	AUC = 0.91		
Padilla et al.*, 2012 [34]	SPECT	Voxel Intensity	FDR NMF	SVM	56 AD 41 CN	91.4	90.6	92.3
Padilla et al.*, 2012 [34]	PET	Voxel Intensity	FDR NMF	SVM	53 AD 41 CN	86.6	87.5	85.4

3

Feature Extraction and Feature Transformation

Contents

3.1 Introduction

The main objective of the current thesis is to build and study a system for the computer-aided diagnosis of Alzheimer's disease, using three-dimensional images produced by the FDG-PET neuroimaging technique. For the purpose of feature extraction, three different approaches were studied. The first uses voxels intensities (VI) which are the features obtained directly from the FDG-PET scan with no further processing. A brief discussion on this type of feature will be presented in the next section. The scale-space of the FDG-PET images was also considered and will be discussed in section 3.3. The other two types of feature, which will be addressed in sections 3.4 and 3.5, are the Local Variance (LVAR) and the Local Binary Patterns (LBPs), respectively.

3.2 Voxel Intensity

Voxel intensity features are obtained directly from the PET scan and its value $V(x, y, z)$ is a direct measure of the FDG uptake detected in a certain voxel. The image database used in the present work, which was retrieved from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [11], had already undergone a preprocessing stage, resulting in a co-registered and normalized set of images with identical dimensions, specifically, 128 by 128 by 60 (section 6.2 provides a detailed description of this preprocessing step). The domain of $V(x, y, z)$, denoted by $\mathcal{B}$, can be stated as follows:

$$\mathcal{B} = \{x, y, z \in \mathbb{N} : 1 \leq x \leq 128, 1 \leq y \leq 128, 1 \leq z \leq 60\}. \tag{3.1}$$

Only one more preprocessing step was carried out on the original images before the feature selection phase. Its aim was to ignore all voxels that lie outside the brain, reducing substantially the dimensionality of the input patterns.

To build a binary mask $M(x, y, z)$, where every position inside the brain is set to true or otherwise set to false, a similar approach to the one used in [33] was utilized. First, an average brain was calculated using the whole VI database and then, the subsequent volume was thresholded at 5% of the maximum value. The threshold was determined empirically so that the brain mask would adapt correctly to the brain. The output of this preprocessing step is illustrated in Figure 3.1.

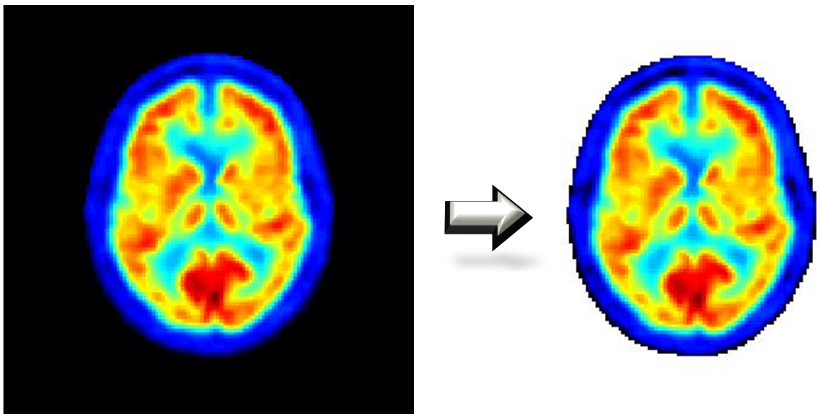


Figure 3.1: Binary mask of the brain. On the left, one example of an axial cut of an arbitrary patient. On the right, the same example but with all voxels outside the brain removed.

Although the procedure just described could be seen as a feature selection routine, it was presented here because it is constructed based on the VI features and should be used as a preprocessing step, not as the only feature selection operation. In addition, the same mask was also used with the other types of feature described in this section.

3.3 Scale-Space Expansion

A common characteristic of images is that neighboring pixels are highly correlated and this remains true for the VI features presented in the previous section. As a consequence, a great deal of information present in the original volume is redundant, which can reduce the system's performance due to the *curse of dimensionality*, as it will be explained in section 4.1. In order to overcome this probable source of performance degradation, a Gaussian pyramid representation of the scale-space of the brain volumes was studied. This pyramid provides representations of an image, in this case of a volume, at different scales and resolutions. In this section, the construction of the scale-space will be reviewed. See [36–38] for more details.

A low-pass pyramid is generated by the repetition of two steps. The first one smooths the volume with an appropriate filter, followed by a subsampling step usually by a factor of two in each direction. More formally, the pyramid is recursively defined as follows:

$$\begin{cases} V_0(x,y,z) = V(x,y,z) \\ V_l(x,y,z) = \displaystyle\sum_{m=-2}^{2}\sum_{n=-2}^{2}\sum_{o=-2}^{2} w(m,n,o)V_{l-1}(2x+m, 2y+n, 2z+o), \quad \text{for} \quad l = 1,2,\ldots, \end{cases} \tag{3.2}$$

where V_l represents the level l of the pyramid and $w(m,n,o)$ is a weighting function, also known as "generating kernel". In this definition of the scale-space, it should be noted that both steps are merged in equation (3.2), the width of the generating kernel was set to five (m, n and o range from -2 up to 2) and the subsampling factor was set to two. A more general definition was not used in order to ease the presentation. On the other hand, the domain of each layer, denoted by $\mathcal{B}_0$, $\mathcal{B}_1$, $\mathcal{B}_2$ and so on, can be defined as follows:

$$\begin{cases} \mathcal{B}_0 = \mathcal{B} \\ \mathcal{B}_l = \left\{x,y,z \in \mathbb{N} : 1 \le x \le \left\lceil \frac{x_{M,l-1}}{2} \right\rceil, 1 \le y \le \left\lceil \frac{y_{M,l-1}}{2} \right\rceil, 1 \le z \le \left\lceil \frac{z_{M,l-1}}{2} \right\rceil \right\}, \end{cases} \tag{3.3}$$

where $l = 1,2,\ldots$, the $\lceil\,\rceil$ notation stands for the ceiling of a number and $x_{M,l-1}$, $y_{M,l-1}$ and $z_{M,l-1}$ are the domain upper limits of x, y and z coordinates in the previous level.

Usually, the generating kernel is constructed so that three properties hold:

- Separability: $w(m,n,o) = w(m) \cdot w(n) \cdot w(o)$;
- Symmetry: $w(-m) = w(m)$, $w(-n) = w(n)$ and $w(-o) = w(o)$;
- Each node at level l should contribute the same total weight to nodes at level $l+1$;

The generating kernel used in work is the one where $w(m) = w(n) = w(o) = \frac{1}{16}\,[1\ 4\ 6\ 4\ 1]$, which resembles the Gaussian function and thus gives rise to the Gaussian pyramid's name. Figure 3.2 shows the generation of the first three levels of the Gaussian pyramid.

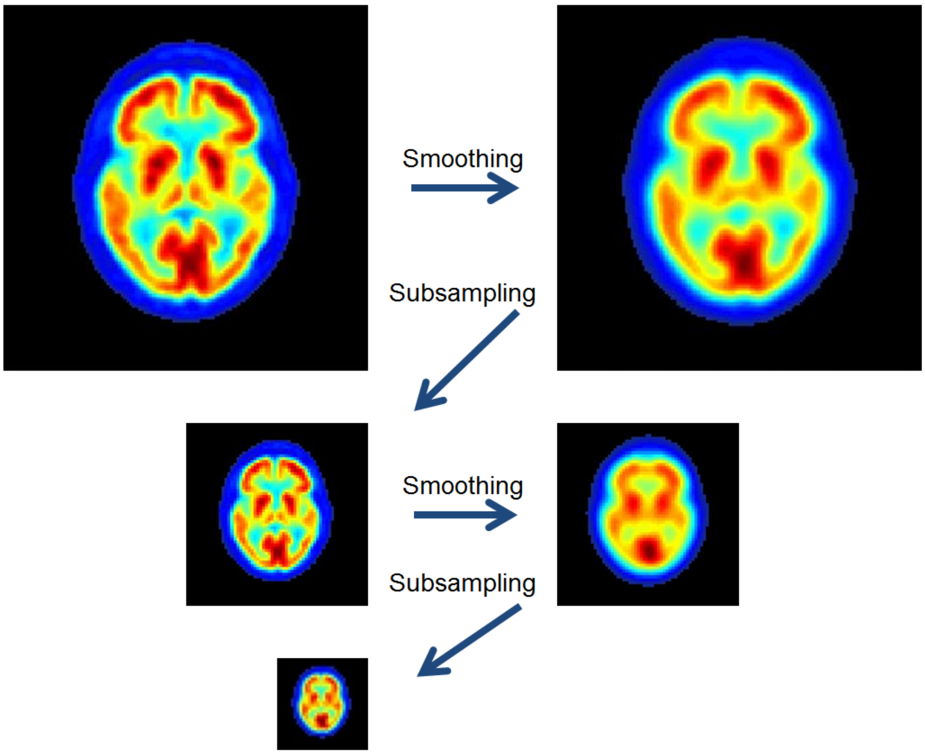


Figure 3.2: Generation of three levels of the Gaussian pyramid, which are illustrated in the images on the left. The images on the right show the output of the intermediate smoothing step. Note that, although only one slice of each brain is depicted, both smoothing and subsampling steps take place in the three-dimensional volume.

3.4 Local Variance

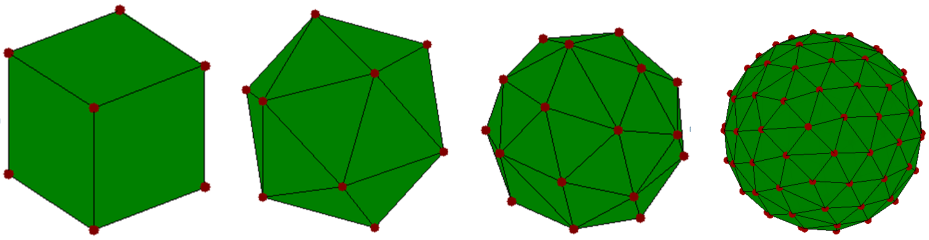
Although VI is the most evident feature to use, since AD is characterized by a diminished brain metabolism and this feature measures that same information, other attributes of the volume produced by the PET scan might also contain discriminative information. In this section, a transformation of the original volume that captures its local contrast will be presented.

The image total variance is one of the many definitions of contrast, known as RMS contrast [39]. However to measure local contrast, one needs to consider the RMS' local counterpart. In fact, areas with low contrast are fundamentally flat, having therefore low variance, while areas near corners or edges have higher contrast and also higher local variance.

In the present work, the 3D nature of the biomarker that is being used for the CAD of AD demands the usage of the variance over a 3D neighborhood, which can be simply defined as the variance of P equidistant sample points $\mathbf{x}_p = (x_p, y_p, z_p)$ with voxel intensities V_p that lie on a sphere with a predefined radius R and centered at a given point $\mathbf{x}_c = (x_c, y_c, z_c)$ (Figure 3.3). This definition of neighbor set has one main advantage: it allows for the extraction of features at different scales by varying the radius R. The operator $VAR_{P,R}$ can therefore be defined as:

$$VAR_{P,R}(\mathbf{x}_c) = \sqrt{\frac{1}{P-1}\sum_{p=1}^{P}\left(V_p(\mathbf{x}_c) - \mu\right)^2}, \quad \text{where } \mu = \frac{1}{P}\sum_{p=1}^{P} V_p(\mathbf{x}_c). \tag{3.4}$$

Hence, if one varies the center $\mathbf{x}_c$, the local contrast of each voxel's neighborhood can be computed.



(a) $P = 8$ **(b)** $P = 12$ **(c)** $P = 24$ **(d)** $P = 98$

Figure 3.3: Neighbor sets for four different numbers of sampling points. Each neighbor point (red) lies on a sphere and is at the same distance to its closest samples. To be more precise, the equidistant property only holds completely accurate for the cases $P = 8$ and $P = 12$, while for the cases $P = 24$ and $P = 98$ an approximation is used.

Since the voxel intensities in use are sampled at specific coordinates on the sphere, i.e., most samples do not belong to the VI domain, $\mathcal{B}$, an interpolated value of V_p must often be calculated. In this case, trilinear interpolation [40] was applied.

Figure 3.4 shows a transformation of an input brain volume based on the operator $VAR_{24,1}$. Note also that one can and should use the binary mask $M(x, y, z)$ described in section 3.2 to reduce the number of features.

Despite the simple formulation of this operator, equidistant sampling on the sphere has no exact solution for most number of sampling points, and the general task is known as *Fejes Toth's* problem. Nevertheless, there are some numerical approximations available, which were used in this work and that can be obtained in [41] and [42]. It is stressed that the exact position of the sampling points is not crucial for this type of feature. However, it will have greater importance for the 3D Local Binary Patterns presented in section 3.5.2.

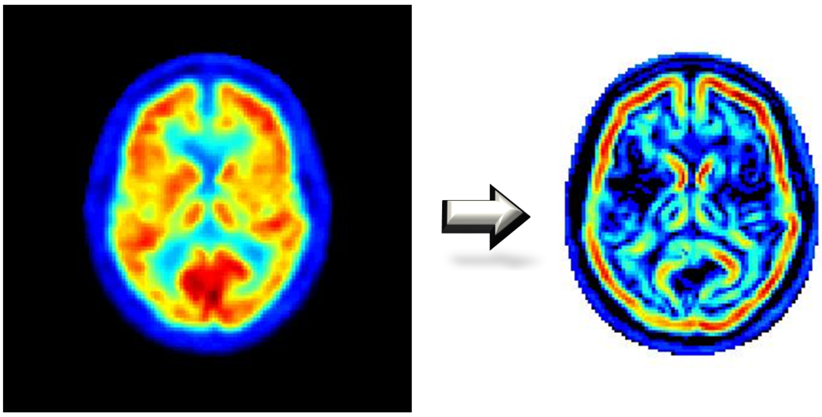


Figure 3.4: Transformation of an input brain volume by the local variance operator based on a neighbor set of 24 samples located on a sphere of radius 1. After the transformation, voxels outside the brain were removed using the brain mask M described in section 3.2. Only one axial cut is depicted for visualization purposes.

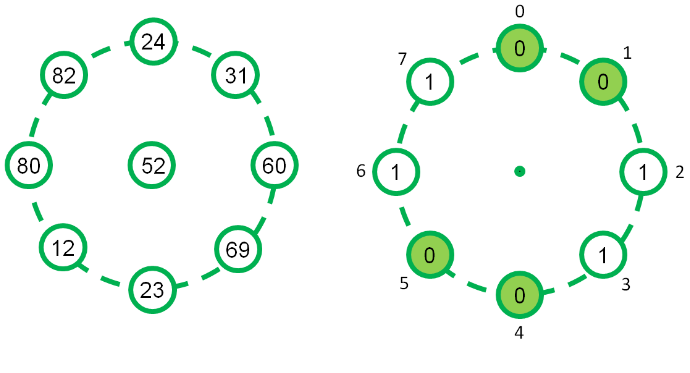
3.5 Local Binary Patterns

Texture is a different source of information that can be retrieved from an image. While the human vision appears to be very efficient in extracting these patterns, its coding for computational purposes is not obvious. As a consequence, there are a great variety of tools to identify texture, such as co-occurance matrices, Laws' texture energy measures, histograms of gradient magnitude and orientation or wavelet coefficients, as can be seen in [43, 44]. One approach that recently has been successfully applied to a wide range of different applications, from texture analysis [45] to face recognition [46], is based on the Local Binary Patterns, LBPs for short.

In the following two sections, these features will be discussed. First of all, in section 3.5.1, the original two-dimensional patterns will be presented, and then a new generalization scheme for the three-dimensional case will be proposed in section 3.5.2. Both 2D and 3D LBPs were implemented and tested in order to evaluate the performance gain achieved with the new scheme.

3.5.1 Two-dimensional LBPs

Local Binary Patterns [47,48] were originally proposed for the analysis of texture in two-dimensional images. An LBP encodes the texture of the local neighborhood of a given pixel $\mathbf{x}_c = (x_c, y_c)$ with gray value V_c, using P equally spaced neighboring pixels with coordinates $\mathbf{x}_p = (x_p, y_p)$ and gray values V_p placed on a circle of radius R. Values at non-integer pixel coordinates are calculated using bilinear interpolation [49]. The encryption is done by thresholding the neighbors with the gray value of the central pixel V_c, yielding a P-dimensional binary vector, $T = \left[H\left(V_1 - V_c\right), H\left(V_2 - V_c\right), \ldots, H\left(V_P - V_c\right)\right]^T$, where $H(\cdot)$ is the Heaviside or unit step function. By assigning a binomial factor 2^p to each term



(a)

(b)

$$\boxed{LBP = 0 \cdot 2^0 + 0 \cdot 2^1 + 1 \cdot 2^2 + \cdots + 1 \cdot 2^7}$$

(c)

Figure 3.5: Computation of a Local Binary Pattern. In **(a)**, all neighbor values are illustrated, together with the central one. The neighbors are then thresholded with the central value in **(b)** and, in **(c)**, the label associated with the pattern is calculated according to equation (3.5).

$H(V_p - V_c)$, one can transform the vector T into a unique $\mathit{LBP}_{R,P}$ label:

$$\mathit{LBP}_{R,P} = \sum_{p=1}^{P} H(V_p - V_c) \cdot 2^p. \tag{3.5}$$

An example can be found in Figure 3.5. Once again, by varying the central pixel $\mathbf{x}_c$, one can associate an LBP label to each position of the image. Afterwards, the histogram of all 2^P LBP labels is computed and used as the texture descriptor. The fact that the number of possible patterns grows exponentially with the number of considered neighbors can pose a problem related to the stability of the histograms. For instance, if 24 neighbors are used, then there are more than 16 million patterns, which means that in order to obtain a stable histogram, a textured image with (a lot) more than 16 million pixels should be used. To alleviate this problem, two distinct extensions were proposed in [48]: uniform LBPs and rotation invariance.

A Local Binary Pattern is said to be uniform if the binary vector T contains at most two transitions from 0 to 1 or vice versa when traversed circularly. This extension is motivated by the fact that uniform patterns have higher incidence rates in textured images. When uniformity is used, uniform LBPs are distinguished with distinct labels and one extra label is used to identify all non-uniform patterns. On the other hand, rotation invariant LBPs do not distinguish between patterns that could be aligned after an appropriate rotation, therefore merging groups of patterns into one label. Rotation invariance in two dimensions can be achieved by assigning to each pattern the label defined by:

$$\mathit{LBP}^{ri}_{R,P} = \min\{\mathit{ROR}\,(T_{R,P}, i) \mid i = 0, 1, \ldots, P-1\}, \tag{3.6}$$

where $T_{R,P}$ is the binary vector T associated with the pattern $\mathit{LBP}_{R,P}$ and $\mathit{ROR}(x, i)$ performs a circular bit-wise right shift on the P-bit number x, i times [48]. Some examples of the two previous extensions are presented in Figure 3.6.

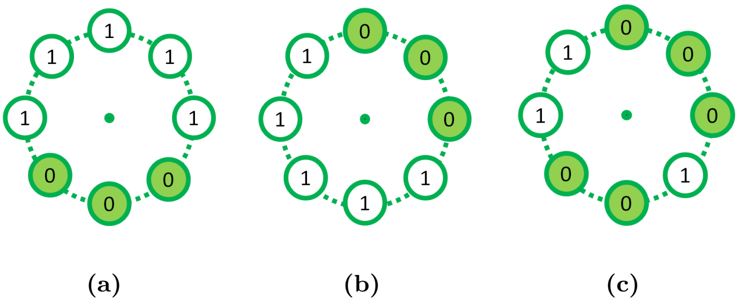
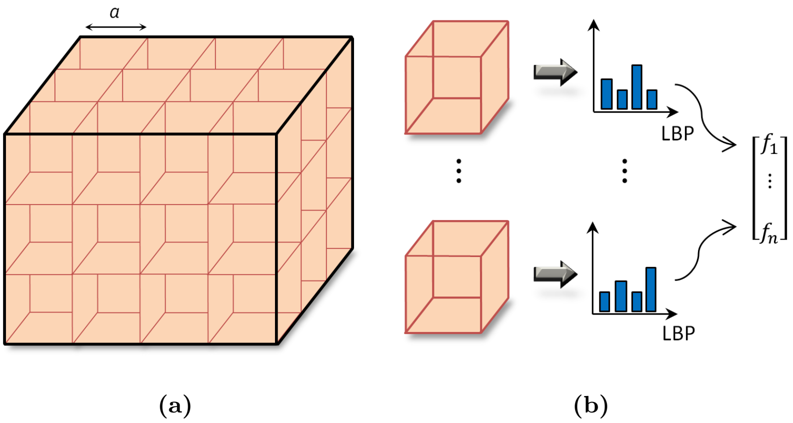


Figure 3.6: Three examples of LBPs. LBPs **(a)** and **(b)** differ only by a rotation and will be merged into label 31, according to equation (3.6), if rotation invariance is considered. LBP **(c)** is different from the previous two and therefore will be labeled differently. On the other hand, LBPs **(a)** and **(b)** are uniform since they only have two transitions from 1 to 0 or vice versa, while **(c)** is non-uniform.

Returning to the problem of the CAD of AD, there are some details regarding the utilization of LBPs that demand special attention. The first one is the fact that the biomarker in use is three-dimensional, while Local Binary Patterns are based on two-dimensional images, reason why a 3D extension is proposed in the next section. Nevertheless, even physicians, while performing diagnosis, analyze only one slice of the brain volume at each time and, thus, one can also use the LBP features computed in each slice. The second problem appears because the brain might not be described by a single texture. In order to solve this problem, the computation of histograms was carried inside cubes of arbitrary dimension, a, in a mesh that spans the entire volume (Figure 3.7(a)), and the resulting texture descriptors were concatenated (Figure 3.7(b)) before proceeding to the feature selection phase. The use of different dimensions of the unit square, a, will allow for the identification of patterns that are present at different scales. The third and last issue concerns the use of rotation invariant patterns. On one hand, the fact that the brain images of all patients already have the same orientation indicates that rotation invariance should not be used in order to achieve the maximum possible discrimination. However, on the other hand, the number of possible patterns is very high and, therefore, the uncertainty associated with the estimation of the probability of occurrence of each pattern, i.e. the histogram, is also considerable, jeopardizing the system's performance. This decision was empirically studied and, at the end, only rotation invariant and uniform patterns were considered.



(a)

(b)

Figure 3.7: The brain volume is partitioned into several disjunct cubes of size a. Then, within each cube, the probability of occurrence of each uniform and rotation invariant LBP is estimated through the use of histograms. The concatenation of all entries of all histograms will form the output of this feature extraction procedure. Adapted from [35].

3.5.2 Three-dimensional LBPs

Several attempts to extend Local Binary Patterns to three-dimensional volumetric data have already been published. However, they all introduce some sort of approximation. In [50], "volume LBPs" are proposed, but they deal with dynamic texture analysis on 2D time series and not full 3D data (as opposed to what the method's name suggests). In [51], not only the neighborhood of a voxel is not thresholded with the central value (the operation is replaced by a simple subtraction) but also a notion of uniformity dependent on the dataset is used, i.e., a given LBP is considered to be uniform if it is one of the most common patterns on that specific dataset. Finally, another approach to rotation invariant LBPs is presented in [52] but the uniformity concept was ignored. In the current work, a novel approach to full three-dimensional, uniform and rotation invariant LBPs is introduced without any approximation of the underlying concepts.

Consider a neighbor set $\{\mathbf{x}_1, \ldots, \mathbf{x}_P\}$ with cardinality P, where all neighbors lie on a sphere with radius R, similar to the one considered for the LVAR type of feature in section 3.4. The gray value of each sample point is denoted by V_p and trilinear interpolation [40] is used to calculate values at non-integer voxel coordinates. The issue of equidistant sampling, which is now necessary for rotation invariance, still remains, but while for some values of P, the exact solution is known, e.g., for 8 sampling points the vertices of a cube lie on a sphere and all points are at the same distance to their closest samples, for other values of P, the approximations given in [41] can be used. The texture of the neighborhood of a given voxel $\mathbf{x}_c = (x_c, y_c, z_c)$ can, therefore, be encrypted in the binary vector $T = [H(V_1 - V_c), H(V_2 - V_c), \ldots, H(V_P - V_c)]^T$ and be labeled with a unique code by assigning a factor 2^p to each term, as it was explained in the previous section. The main obstacles arise when one tries to introduce uniformity and rotation invariance to reduce the number of distinct LBPs.

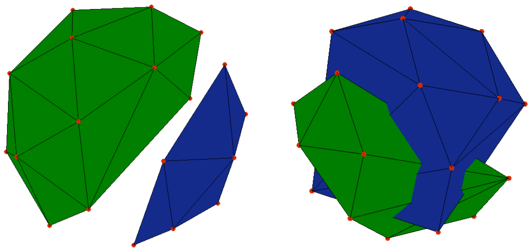
First, a different definition of uniformity is necessary, so that it can be generalized to higher dimensions. The following definition is proposed: an LBP is considered to be uniform if and only if the convex hull $\mathcal{H}_0$ of the neighbor points where $H(V_p - V_c) = 0$, and the convex hull $\mathcal{H}_1$ of the remaining ones do not intersect, as illustrated in Figure 3.8. Note this definition can be applied directly to the original 2D situation, leading to the same notion of uniformity. Now, since the convex hull of a set of points is known to be a polyhedron, one can represent $\mathcal{H}_i$ by a system of m_i linear inequalities, which in matrix form is given by:

$$\mathcal{H}_i : \quad A_i x \leq b_i, \tag{3.7}$$

where $A_i \in \mathbb{R}^{m_i \times D}$, $x \in \mathbb{R}^D$, $b_i \in \mathbb{R}^{m_i}$ and D is the number of spatial dimensions (2 or 3). The intersection $\mathcal{I}$ is, therefore, simply given by the following system of $m = m_0 + m_1$ inequalities:

$$\mathcal{I} = \mathcal{H}_0 \cap \mathcal{H}_1 : \quad \begin{bmatrix} A_0 \\ A_1 \end{bmatrix} x \leq \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}. \tag{3.8}$$

The feasibility/unfeasibility of the previous system was determined using the *B-rule* algorithm proposed in [53], which either finds a solution x to the linear system or it gives a conclusive proof that no such vector x exists. Since the exposition of the algorithm, although not complex, cannot be done in just a few paragraphs, the interested reader should be reported to the original paper for more details.



(a) Uniform LBP **(b)** Non-uniform LBP

Figure 3.8: Two distinct 3D-LBPs, one uniform **(a)** and one non-uniform **(b)**, are depicted. For both, the green convex hull is formed by the points where $H(V_p - V_c) = 1$, and the blue by the points where $H(V_p - V_c) = 0$. The intersection of the convex hulls dictates the uniformity of the pattern.

After identifying all uniform patterns one by one, rotation invariance should be taken into account. The goal is to merge patterns that differ only by a rotation, without having to explicitly query against all possible transformations. This task can be accomplished using spherical harmonics, as it will be described below. A brief review containing the basic concepts and important properties of spherical harmonics that proved to be useful for this thesis is given in Appendix A.

The following procedure was carried out to detect rotation invariant LBPs. First, a square-integrable spherical function $f(\theta,\varphi)$ was defined with value one in a small neighborhood of area A of every point $\mathbf{x}_p$ where $H(V_p - V_c) = 1$, and zero everywhere else, as shown in Figure 3.9.

$$f(\theta,\varphi) = \begin{cases} H(V_p - V_c) & , \; ||\mathbf{x} - \mathbf{x}_p||^2 \leq \varepsilon \quad \text{for} \quad p = 1, \ldots, P \\ 0 & , \; \text{otherwise} \end{cases} \tag{3.9}$$

In the previous equation, $\mathbf{x}$ is restricted to the sphere. Then, the function $f(\theta,\varphi)$, characteristic of a given LBP, was decomposed into its harmonics:

$$f(\theta,\varphi) \approx \sum_{l=0}^{l_M} \sum_{m=-l}^{l} a_{l,m} Y_l^m(\theta,\varphi), \tag{3.10}$$

where $a_{l,m}$ is the complex coefficient associated with the spherical harmonic Y_l^m of degree l and order m, and l_M is the maximum degree of expansion which was set, empirically, high enough so that different patterns could be correctly distinguished. The harmonic coefficients were computed using equation (A.4), and by noting that as the area A tends to zero, the function $Y_l^m(\theta,\varphi)$ restricted to that small region becomes approximately constant, which means that:

$$a_{l,m} = \int_0^{2\pi}\int_0^{\pi} f(\theta,\varphi) \cdot Y_l^{m*}(\theta,\varphi) d\Omega \approx \sum_{p=1}^{P} H(V_p - V_c) \cdot Y_l^{m*}(\theta_p,\varphi_p) \cdot A, \tag{3.11}$$

where (θ_p,φ_p) is the location of the p-th neighbor in spherical coordinates and the star notation stands for the complex conjugate of a function.

The actual value of A is not significant, as it will be seen shortly, and thus can be set arbitrarily small, but not zero, so that the previous equality holds true in the limit. Finally, the rotation invariant shape descriptor SD, defined in equation (3.12), of the function f was used to uniquely identify a rotation invariant LBP.

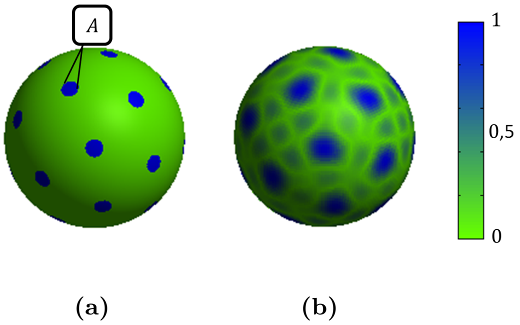


Figure 3.9: Shape corresponding to the LBP where all 24 neighbor intensities are greater than the intensity of the central voxel, i.e., $H(V_p - V_c) = 1$ for all p. **(a)** – Exact function. **(b)** – Reconstruction using spherical harmonics. The reconstruction was accomplished with a small maximum degree of expansion, but as this parameter grows, the reconstruction will approach the original function.

$$\text{SD} = \left\{||a_{0,0}|| \ ; \ ||(a_{1,-1}, a_{1,0}, a_{1,1})|| \ ; \ \dots \ ; \ ||(a_{l_M,-l_M}, \dots, a_{l_M,l_M})||\right\} \tag{3.12}$$

It should be noted that when the harmonic coefficients given in equation (3.11) are substituted into the descriptor (3.12), the factor A is common to all elements and, therefore, can be eliminated without losing any discriminative power. In order to prove that SD is, in fact, invariant under rotations, consider first the descriptor proposed by Michael Kazhdan et al. in [54] and defined by:

$$\text{SH}(f) = \{||\pi_0\,(f)\,|| \ ; \ ||\pi_1\,(f)\,|| \ ; \ \dots \ ; \ ||\pi_{l_M}\,(f)\,||\}\,, \tag{3.13}$$

where π_l, which is defined by equation (A.7), represents the projection into the subspace formed by the span of the spherical harmonics with l fixed to a given degree.

The descriptor SH is invariant under rotations, as demonstrated in [54] and reproduced below.

$$\begin{aligned}
\text{SH}(R(f)) &= \{||\pi_0\,(R\,(f))\,|| \ ; \ ||\pi_1\,(R\,(f))\,|| \ ; \ \dots \ ; \ ||\pi_{l_M}\,(R\,(f))\,||\} && (3.14)\\
&= \left\{\left|\left|R\,(\pi_0\,(f))\right|\right| \ ; \ \left|\left|R\,(\pi_1\,(f))\right|\right| \ ; \ \dots \ ; \ \left|\left|R\,(\pi_{l_M}\,(f))\right|\right|\right\} && (3.15)\\
&= \left\{\left|\left|\pi_0\,(f)\right|\right| \ ; \ \left|\left|\pi_1\,(f)\right|\right| \ ; \ \dots \ ; \ \left|\left|\pi_{l_M}\,(f)\right|\right|\right\} && (3.16)\\
&= \text{SH}(f) && (3.17)
\end{aligned}$$

Note that step (3.15) arises from property (A.6) which simply states that the projections π_l commute with rotations, and step (3.16) from the fact that rotations do not change the norm of a function. Finally, using the orthonormality of the spherical harmonics base functions, one can conclude that SD and SH represent the same descriptor, therefore also proving the rotation invariance of the former.

$$\begin{aligned}
\left|\left|\pi_l(f)\right|\right|^2 &= \left|\left|\textstyle\sum_{m=-l}^{l} a_{lm} Y_l^m\right|\right|^2 && (3.18)\\
&= \iint_\Omega \left(\textstyle\sum_{m=-l}^{l} a_{lm} Y_l^m\right)\left(\textstyle\sum_{m=-l}^{l} a_{lm} Y_l^m\right)^* d\Omega && (3.19)\\
&= \iint_\Omega \textstyle\sum_{m=-l}^{l} a_{lm} a_{lm}^* \cdot Y_l^m Y_l^{m*} d\Omega && (3.20)\\
&= \textstyle\sum_{m=-l}^{l} a_{lm} a_{lm}^* \cdot \iint_\Omega Y_l^m Y_l^{m*} d\Omega && (3.21)\\
&= \textstyle\sum_{m=-l}^{l} a_{lm} a_{lm}^* && (3.22)\\
&= \left|\left|(a_{l-l}, \dots, a_{ll})\right|\right|^2 && (3.23)
\end{aligned}$$

In the previous chain of equalities, step (3.18) comes from equation (A.7), step (3.20) uses the orthogonality of the base functions Y_l^m and step (3.22) the normality.

On a different note, since for some cardinalities of the neighbor set, the equidistant sampling is only an approximation, thus affecting rotation invariance, a small difference between the SD descriptors was allowed. More precisely, if one thinks of SD as a vector of dimension $l_M + 1$, the same label is assigned to two LBPs with descriptors SD_i and SD_j if:

$$\frac{||\text{SD}_i - \text{SD}_j||}{\max\{||\text{SD}_i||, ||\text{SD}_j||\}} \le \eta, \tag{3.24}$$

and if a given pattern lies within this margin with two distinctly labeled LBPs, then the first is assigned to the group of the closest pattern. The closeness criterion was defined as in the left-hand side of inequality (3.24). In addition, parameter η should be small in order to allow only small differences in rotation invariant patterns. Bearing this in mind, different values of η were studied experimentally, and the parameter was fixed at 0.05 in the end.

It is now possible to build a map that links every 2^P possible patterns to uniform and rotation invariant LBP labels. First, all patterns have to be classified as uniform/non-uniform LBPs and a unique label is given to all non-uniform. This step imposes a computational limit on the number of neighbors in use, since its time complexity grows exponentially with P. Second, the SD descriptor has to be built for all uniform LBPs and then, while the ones with the same descriptor are tagged with the same label, different labels are given to LBPs with different descriptors. The final step of the current extraction procedure is to label each position of each subject's brain image using the mapping previously constructed and, then, compute several histograms, each one in a different region, as illustrated in Figure 3.7. The resultant feature vector is formed by the concatenation of all histogram entries, where each entry measures the incidence rate of a given uniform and rotation invariant LBP.

3.6 Conclusion

Three distinct features were covered in the current chapter: VI, LVAR and LBP.

Regarding the VI features from the FDG-PET scan, the scale-space of the brain images was studied, allowing for a reduction of the number of features. Note that the number of voxels is reduced by a factor of eight (two in each space direction) in each level. The dimensionality reduction achieved by the pyramid representation of the scale-space has three main objectives. First, there is the possibility of improving the system's performance by alleviating the small sample size problem. Second, it reduces the time consumed at the training stage and third, it allows studying how much data could be discarded without jeopardizing the system's performance.

A measure of local contrast, LVAR, was also introduced as the sample variance computed on a given 3D neighborhood. Since this type of feature estimates the variance of the image intensity on a given sphere for each position of the PET scan, the number of neighbors P has to be set high enough so that good estimates of the true variance can be computed.

Finally, the feature LBP, which is a texture descriptor, was presented starting by the key concepts originally proposed for 2D textured images, and then proposing a novel generalization to three dimensions. The resulting features are the entries of several histograms, where each entry is associated with the occurrence of each distinctly labeled LBP and each histogram with a different region of the brain. It was also shown that rotation invariance and uniformity are fundamental to reduce the number of LBP labels, which is important to achieve stable histograms, i.e., histograms that represent a good estimation of the probability of occurrence of each label. On a different note, the new generalization proposed for three-dimensional data does not introduce any approximation to the original concepts. It is however limited by the exponential growth of the number of possible patterns with the number of neighbors considered. Nevertheless, classification results (see section 6.5) showed that good generalization ability can still be achieved with a low cardinality of the neighbor set.

4 Feature Selection

Contents

4.1 Introduction

4.1.1 Motivation

The feature extraction methods presented in the previous chapter aim to create features that are discriminative to the classification task at hand, but they do not account for the statistical value of the information that can be extracted from the dataset. Another problem related to the feature extraction methods exploited in this study is that the number of features produced can easily be as high as tens or even hundreds of thousands.

It is known that such high dimensionality combined with a comparatively small sample size usually leads to a degradation of the classifier's performance, a phenomenon known as the *curse of dimensionality*, broadly speaking, because with more dimensions it becomes easier to overfit, i.e., to find accidental regularities in the training set, not present in different unseen data, and therefore leading to poorer generalization ability. In theory, the probability of misclassification of a given decision rule does not increase with the number of used features, as long as the probability mass functions conditioned on the class are known. However, since in such high dimensionality cases that is not possible, due to the limited sample size, in practice the performance accuracy may actually be degraded by the added features [55]. In fact, when a generative approach to the learning stage is undertaken, classification principles are based on parametric estimates of the class-conditional densities. For a fixed sample size, as the number of features increases (with the corresponding increase in the number of unknown parameters), the reliability of the parameters estimates decreases and, consequently, the performance of the resulting classifier may deteriorate [55–57]. By contrast, discriminative models are widely regarded as less complex than generative models and, as a result, they are more robust to the *curse of dimensionality* and demand smaller training sets [58]. Nevertheless, even when using a discriminative approach, a well-defined dimensionality reduction scheme might improve the performance of any pattern recognition system.

Feature selection procedures, which as the name suggests reduce the dimensionality of the input vectors by selecting only a subset of features, achieve other important objectives, alongside with the improvement of the system's performance already discussed: both learning and classification steps are speeded up, measurement and storage requirements are reduced and data visualization and data understanding is facilitated [59].

4.1.2 Problem Formalization and Notation

Formally, the feature subset selection problem can be posed in the following way. Let $\boldsymbol{S}$ be the input dataset formed by K samples:

$$\boldsymbol{S} = \left\{ \left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(K)}, y^{(K)}\right) \right\}, \tag{4.1}$$

each one consisting of D input variables $\mathbf{x}^{(k)} = \left(x_1^{(k)}, \ldots, x_D^{(k)}\right)$ produced by some feature extraction procedure, such as the ones discussed in the previous chapter, and one output variable or class label $y^{(k)}$. The goal of feature selection is to find a subspace of N features $\mathbb{R}^N$ from the D-dimensional observation space $\mathbb{R}^D$ that "optimally" describes the class label. The ultimate criterion would be the minimization of the Bayes error of predicting the class label from the subset of chosen variables. Note, however, that this criterion cannot be used in practice because the true probability distribution that models the data is generally not known, and thus, other metrics must be applied.

In order to completely specify the notation used throughout this chapter (except in section 4.4 where a different notation will be established because medically driven selection procedures are not based on the training set $\boldsymbol{S}$), additional notation is now introduced. In situations where the input vector $\mathbf{x}$ can be interpreted as the realization of a random variable, the random variable that models the i-th component of $\mathbf{x}$ will be denoted by X_i, and the set of all random variables X_i by $\boldsymbol{X}$. Similarly, Y will be the random variable of which each $y^{(k)}$ is a realization. Additionally, the K-dimensional vector containing all realizations of the i-th variable will be denoted by $\mathbf{x}_i$ and the vector containing all K class labels by $\mathbf{y}$.

On a different note, feature selection methods are usually categorized into two different classes: *Wrappers* that assess subsets of features according to their usefulness to a given predictor, or in other words, according to classification results obtained by a given classifier trained with each subset of features, and *filters* which select subsets of variables as a preprocessing step, independently of the chosen predictor [60]. The main advantage of the first class of methods is that they achieve better performances, since they optimize the classification performance directly. However, compared to *filters*, *wrappers* are substantially slower, especially in higher dimensions as it is the case, reason why only *filter* methods were considered. Specifically, one procedure based on the Pearson correlation coefficient which will be described in section 4.2, and two methods based on mutual information, Mutual Information Maximization (MIM) and Minimal Redundancy Maximal Relevance (mRMR), which will be described in sections 4.3.1 and 4.3.2, respectively. In this thesis, an additional category of selection procedures based on medical information was explored. This class of selection procedures uses the expertise of a physician to select the best features, instead of using the information contained in the training instances. Two techniques based on the information of the movement of a trained physician's eye will be presented, namely Time Independent Eye Track Driven Selection (TI-ETDS) in section 4.4.2 and Time Dependent Eye Track Driven Selection (TD-ETDS) in section 4.4.3.

4.2 Feature Selection based on Correlation Coefficients

Correlation coefficients, such as Pearson correlation coefficient, measure the amount of correlation (linear dependence) between two variables [61]. Therefore, the utility of the i-th feature X_i can be quantified by the Pearson correlation coefficient between the feature itself and the class label Y, which is defined by:

$$\mathcal{R}(X_i, Y) = \frac{cov(X_i, Y)}{\sqrt{var(X_i)var(Y)}}, \tag{4.2}$$

where *cov* stands for covariance and *var* for variance. An estimator of $\mathcal{R}(X_i, Y)$ is given by:

$$R(X_i, Y) = \frac{\sum_{k=1}^{K} \left(x_i^{(k)} - \bar{\mathbf{x}}_i\right)\left(y^{(k)} - \bar{\mathbf{y}}\right)}{\sqrt{\sum_{k=1}^{K} \left(x_i^{(k)} - \bar{\mathbf{x}}_i\right)^2 \sum_{k=1}^{K} \left(y^{(k)} - \bar{\mathbf{y}}\right)^2}}, \tag{4.3}$$

where the bar notation designates the average over all samples. In the classification problem at hand, the output variable Y is dichotomous, and thus if only the values 0 and 1 are allowed for the class label, equation (4.3) can be simplified into:

$$R_{pb}(X_i, Y) = \frac{\mu_1 - \mu_0}{s_{X_i}} \sqrt{\frac{k_1 k_0}{K^2}}, \tag{4.4}$$

where μ_1 and μ_0 are the mean values of the variable X_i, when only samples with class label 1 and 0 are used, respectively. k_1 and k_0 are the number of samples within each class and s_{X_i} is the estimate of the standard deviation of X_i, given by:

$$s_{X_i} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(x_i^{(k)} - \bar{\mathbf{x}}_i \right)^2}. \tag{4.5}$$

In linear regression, R^2 represents the fraction of the total variance of one variable that can be explained by the other using a linear predictor, and thus, if $R(X_i, Y)^2$ is used as a feature ranking criterion, features are selected according to their individual goodness of linear fit. In addition, the correlation coefficient ranges from -1 to 1, with the extreme values implying a perfect linear dependency between a given feature and the output variable, which means that $R(X_i, Y)^2$ ranges from 0 to 1, with values close to 1 being good indicators for the feature's relevance.

To conclude, the feature selection method based on the Point Biserial Correlation Coefficient (PBCC) can now be fully stated. First, it computes R^2_{pb}, using equation (4.4), for each feature X_i present in the starting feature set $\boldsymbol{X}$ and, after sorting all correlation coefficients, selects the top N features.

4.3 Feature Selection based on Mutual Information

One of the main disadvantages of correlation coefficients is that they only take into account linear dependencies between a given feature and the class label. A better measure of information dependency arises from information theory and is known as mutual information. Given two random variables W and Z, their mutual information is defined as the Kullback-Leibler divergence of the product of their marginal distributions $p(w)p(z)$ from the random variables' joint distribution $p(w, z)$ [62]:

$$I(W; Z) = \iint p(w, z) \log \frac{p(w, z)}{p(w)p(z)} dwdz, \tag{4.6}$$

or equivalently, in a discrete case:

$$I(W; Z) = \sum_{w \in \mathcal{W}} \sum_{z \in \mathcal{Z}} P(w, z) \log \frac{P(w, z)}{P(w)P(z)}, \tag{4.7}$$

where $\mathcal{W}$ and $\mathcal{Z}$ are the dictionaries containing all possible events of the random variables W and Z, respectively. It is worth noting that when w and z are independent from each other, which means that no information about one variable can be extracted from the other, $p(w, z)$ becomes $p(w)p(z)$ and $I(W, Z)$ is reduced to zero.

One of the most compelling arguments supporting the use of mutual information in feature selection is given by *Fano's Inequality* which imposes a lower bound on the Bayes probability error P_e [63]:

$$P_e = P(g(X) \neq Y) \geq \frac{H(Y) - I(X; Y) - 1}{\log(|Y|)}, \tag{4.8}$$

stating that the probability of misclassification is lower bounded by an expression dependent on the mutual information, regardless of the decision function $g(X)$, where X represents a subset of the set containing all starting features $\boldsymbol{X}$. As the mutual information grows, the bound comes closer to zero, but whether or not the bound is actually reached, depends only on the ability of the classifier $g(X)$. However, there are also bad news. X is usually a large subset of random variables and thus, estimating the joint mutual information $I(X;Y)$ involves the estimation of high dimensional density functions which is extremely difficult and unreliable due to the sparsity of the data.

In a recent unifying work, Gavin Brown [64] developed a framework that is able to subsume almost every feature selection method based on mutual information that is currently known, and therefore will be used to discuss both methods used in this study. First, he expanded the joint mutual information $I(X;Y)$ into a sum of interaction information terms:

$$I(X;Y) = \sum_{T \subseteq X} I(\{T \cup Y\}), \quad |T| \geq 1, \tag{4.9}$$

where $\sum_{T \subseteq X}$ should be read as the sum over all possible subsets T drawn from X. Then, he truncated the expansion, ignoring information terms that take into account more than 2 features ($|T| > 2$) and the class label, yielding:

$$I(X;Y) \approx \sum_{i=1}^{N'} I(X_i;Y) + \sum_{i=1}^{N'} \sum_{j=i+1}^{N'} I(X_i;X_j;Y), \tag{4.10}$$

where N' is the cardinality of the set X.

Finally, he considered an incremental search algorithm, or in other words, an iterative algorithm that selects one feature in each step. Thus, if $n-1$ features have already been chosen, the utility of including the n-th feature X_n is quantified by $I(X_n;Y|X_{1:n-1}) = I(X_{1:n};Y) - I(X_{1:n-1};Y)$. Using equation (4.10), an estimator for this information gain is given by:

$$I(X_n;Y|X_{1:n-1}) \approx I(X_n;Y) + \sum_{k=1}^{n-1} I(X_n;X_k;Y), \tag{4.11}$$

which, using the definition of interaction information $\big(I(A;B;C) = I(A;B|C) - I(A;B)\big)$, can be rewritten as:

$$I(X_n;Y|X_{1:n-1}) \approx I(X_n;Y) - \sum_{k=1}^{n-1} \big[I(X_n;X_k) - I(X_n;X_k|Y)\big]. \tag{4.12}$$

The First Order Utility (FOU) for the n-th feature X_n, as the utility criterion (4.12) was called by the author, is composed of three parts: its own mutual information with the target variable, a second term that penalizes redundant features when compared with all features already selected and a third positive term that takes into account class-conditional correlations. This term, which is often forgotten in the literature, states that the inclusion of correlated features might improve the discriminative power of the set of selected features as long as the correlation within each class is stronger than the overall correlation [65]. An illustrative example of the importance of this term is given in Figure 4.1 and explained in its caption. Finally, Brown realized that if he introduced a weighting coefficient in each term, as can be seen in equation (4.13), then almost every feature selection method in the literature is subsumed.

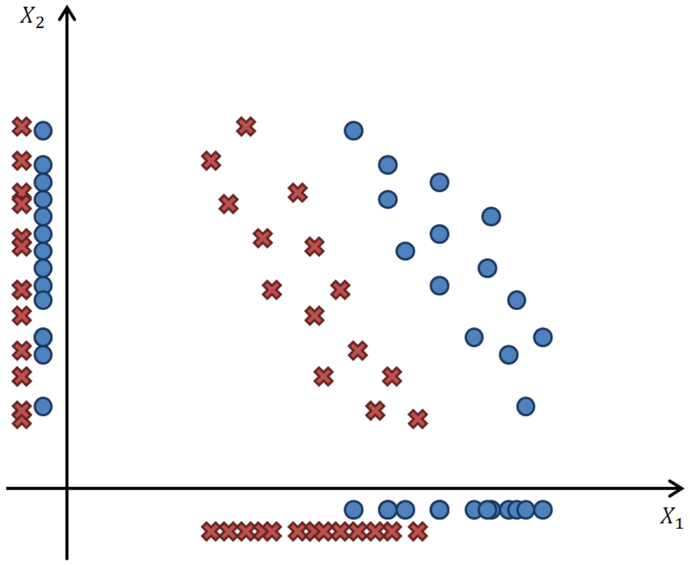


Figure 4.1: The role of class conditional mutual information in feature selection. In this classification task, feature X_1 contains some information about the class and feature X_2 is completely useless (when considered separately). Nevertheless, they complement each other and together they completely separate the data. Consider now a situation where the only selected feature is X_1. Consequently, if the class conditional correlation term is ignored, the inclusion of X_2 will appear unhelpful since, not only it is poorly discriminative (low $I(X_2; Y)$), but also it is (anti-)correlated with X_1 (high $I(X_1; X_2)$). On the contrary, if the conditional correlation is considered, then one can conclude that X_1 and X_2 together are highly discriminative. To see why, note that all three terms of equation (4.12) ($I(X_2; Y) - I(X_2; X_1) + I(X_2; X_1|Y)$) can be written as the conditional mutual information $I(X_2; Y|X_1)$ which is able to capture their joint predictive power.

$$J(X_n) = I(X_n; Y) - \beta \sum_{k=1}^{n-1} I(X_n; X_k) + \gamma \sum_{k=1}^{n-1} I(X_n; X_k|Y) \tag{4.13}$$

In the present work, two different methods based on Mutual Information are studied and can also be described within this framework. They will be presented in the next two sections.

4.3.1 Mutual Information Maximization

The first method is known as Mutual Information Maximization (MIM). It is a ranking algorithm, very similar to the one based on correlation coefficients, but instead of using the PBCC as the ranking criterion, it uses the mutual information between the feature itself and the class label.

$$J(X_i) = I(X_i; Y), \qquad X_i \in \boldsymbol{X} \tag{4.14}$$

Within the framework described in the previous section, this method corresponds to the simplest setting with both coefficients, β and γ, being zero. This realization provides a hint to the two principal setbacks of this method. First, it does not consider the redundancy between features ($\beta = 0$) and thus if two features are highly discriminative but highly correlated, e.g. one is simply the repetition of the other, MIM will select both, despite the same information could be accounted for with only one of them. On the other hand, MIM does not consider the mutual information between pairs of features conditioned on the class ($\gamma = 0$). Note also that, although Brown's framework is based on an incremental search, since no interactions between pairs of features are taken into account, $\beta = \gamma = 0$, then the inclusion of a given feature will not change the utility $J(X_i)$ of the others, and thus it is possible to treat this method as a ranking algorithm instead of an incremental one.

On a different subject, a common approach to the estimation of the mutual information $I(X_i;Y)$, and the one implemented is this study, uses histograms to estimate both marginal and joint density functions and, then, utilizes the definition of mutual information given in equation (4.7). This approach also demotes continuous random variables to discrete by partitioning the space in equal segments, and estimates each probability by counting the number of elements in each partition. The main advantage of histograms is their computational efficiency, when compared with other non-parametric probability density estimators such as Parzen-Window estimators.

4.3.2 Minimal Redundancy Maximal Relevance

The second algorithm, Minimal Redundancy Maximal Relevance (mRMR) [24], is incremental and includes the second term of equation (4.13) into the feature utility function. This second term is used to avoid selecting features that are redundant, despite having high discriminative power when considered individually.

Formally, mRMR can be described as follows. Consider two sets of features: the set $\boldsymbol{D}_t$ containing all features selected at time t and the set $\boldsymbol{F}_t$ with the remaining ones, such that the equality $\{\boldsymbol{D}_t \cup \boldsymbol{F}_t\} = \boldsymbol{X}$ holds. Initially, the set $\boldsymbol{D}_0$ is empty and the set $\boldsymbol{F}_0$ contains all features. Then, at each time step t, mRMR selects from $\boldsymbol{F}_t$ the feature that maximizes the utility function:

$$J(X_i) = I(X_i;Y) - \frac{1}{|\boldsymbol{D}_t|} \sum_{X_j \in \boldsymbol{D}_t} I(X_i;X_j), \qquad X_i \in \boldsymbol{F}_t. \tag{4.15}$$

The selected feature is removed from the set $\boldsymbol{F}_t$ and added to $\boldsymbol{D}_t$ and the same procedure is repeated until the desired number of features, N, is reached. As can be seen, the parameter β of equation (4.13) is used to average the mutual information between a given feature and all previously selected ones, $\beta = 1/|\boldsymbol{D}_t|$.

It was considered that the inclusion of this second term would be important in the context of the CAD of AD because it is known beforehand that features like voxel intensity or local variance have rich redundancy when one considers nearby voxels, and thus selecting more than one voxel from the same neighborhood will probably not change the class-discriminative power of the selected subset much.

Once again, the estimation of the probability mass functions, necessary for the computation of the mutual information, was accomplished through frequency counts. In this selection algorithm, the computational efficiency of histograms is even more important, since mRMR is computationally heavier than the ranking methods presented before.

On a different note, no method that takes into account the third term of equation (4.13) was considered in this study because it was already difficult to get sufficient amounts of results with mRMR and including this term (for example, with the method proposed by Yang and Moody – Joint Mutual Information [66] – or the one proposed by François Fleuret – Conditional Mutual Information [67]) would result in an even greater slowdown.

4.4 Eye Track Driven Selection

The ability to process and decide upon the tremendous amounts of information continuously received by the human brain is a result of several million years of biological evolution. As a result, the human mind is able to recognize almost instantly a great variety of patterns in a process that even the subject himself is not fully aware. One important example is the human vision and the way the human brain processes visual data. Human vision is a piecemeal process relying on the perceptual integration of small regions to construct a coherent representation of the whole [68]. In this process, the mobility of the eyes is used to scan the scene that surrounds the individual and to focus on regions of interest, bringing them into high resolution [68].

Nowadays, the eye tracking technology can record at each time the point of gaze in order to find which regions captured most of a person's attention, providing therefore a way to understand how the subject perceived the scene he or she was seeing [68]. This technology has been recently used in a variety of applications, for instance in marketing research or in web site assessment. In this study, it will be used to capture the medical expertise associated with the examination of FDG-PET images.

4.4.1 Physician Eye Tracking

Eye tracking data used in the present work was acquired during an experiment led by Bicacro et al., reported in [18] and summarized in this section. The data associated with the movement of the physician's eye was recorded using the T120® model from Tobii™ [69]. This model is able to measure the gaze point on a flat screen with an accuracy of 0.5 degrees through an advanced form of the Pupil Centre Corneal Reflection technique. The Tobii eye tracker generates reflection patterns on the corneas of the user's eyes using near infra-red light, which are then collected by image sensors, together with other visual information about the person. Finally, all this information is combined to calculate the three-dimensional position of each eyeball, and eventually the two-dimensional gaze point on the screen [70].

An expert physician was asked to examine all FDG-PET scans from the dataset utilized in this work. The T120® eye tracker was then used to record the path created by the movement of the physician's eyes, while he was examining each of the patients' scans.

The final output of Bicacro's experiment, and an input to this work, was n_k time-dependent sequences of positions $\boldsymbol{X}_{t,s}^{(k)} = (x, y, z)_{t,s}^{(k)}$ focused by the physician for each patient k, with each sequence s restricted to a specific slice z, together with the total amount of time $d_{t,s}^{(k)}$ spent in each location:

$$\left\{\{(\boldsymbol{X}, d)_t\}_s\right\}^{(k)} \qquad \text{for} \qquad t \in \{1, \ldots, T_s^{(k)}\},\ s \in \{1, \ldots, n_k\}\,,\ k \in \{1, \ldots, K\}\,, \tag{4.16}$$

where $T_s^{(k)}$ is the number of gazed points in the sequence s for the patient k. Note that the physician can only analyze one axial cut of the three-dimensional image at a time, reason why the coordinate z remains constant within each sequence.

This information was used to build a probabilistic model of the voxels that caught the physician's attention, in order to be able to select, randomly from that model, the desired number of features. In the next two sections, two different models will be described: one that ignores the sequence through which the physician gazed the brain images and another that takes this information into account.

4.4.2 Time-Independent Eye Track Driven Selection

The amount of time spent by a physician analyzing a given location in the brain volume may be a good indicator of the usefulness of the voxels present therein. Bearing this in mind, one can use the probability $P(\mathbf{x})$ of a given voxel $\mathbf{x} = (x, y, z)$ being used by the physician during a diagnosis to randomly select N voxels and, then, use their intensities as features.

The estimation of the probability function $P(\mathbf{x})$ was accomplished using Parzen-Windows [71] with a Gaussian kernel (see Appendix B for a brief revision), where every point $\boldsymbol{X}_{t,s}^{(k)}$ associated with every person in the training set, regardless of its instant t and sequence s, was used as a sample with weight $d_{t,s}^{(k)}$ or, to be in consensus with equation (B.1) where no weights are considered, each point was used as $d_{t,s}^{(k)}$ samples, i.e.:

$$\hat{P}(\mathbf{x}) = \frac{1}{\sum_{k,s,t} d_{t,s}^{(k)}} \cdot \sum_{k,s,t} d_{t,s}^{(k)} \phi(\mathbf{x} - \mathbf{X}_{t,s}^{(k)}, h), \tag{4.17}$$

where ϕ is the Gaussian kernel defined in equation (B.3) and h is the kernel width.

This approach was preferred over the simpler methods based on histograms for two main reasons. Firstly, the amount of data available from the eye tracking experiment is very small when compared with the total number of possible events (number of voxels in the brain volume) and, thus, approaches based on frequency counts would yield unstable estimators. Secondly, the assumption that the physician was extracting information from a single voxel when analyzing a particular point is too restrictive, since there is a visual angle in which visual acuity is maximal, and therefore neighboring points to the one fixated should also be weighted in the density function as it is done by the Gaussian kernel of the Parzen-Window estimator.

In practice, the Parzen-Window estimate $\hat{P}(\mathbf{x})$ can be obtained by the discrete convolution between the kernel $\phi(\mathbf{x})$ restricted to integer coordinates, and an image $f(\mathbf{x})$ formed by the weighted sum of discrete unit impulses $\delta(\cdot)$, each one centered on a different sample, i.e.:

$$f(\mathbf{x}) = \frac{1}{\sum_{k,s,t} d_{t,s}^{(k)}} \cdot \sum_{k,s,t} d_{t,s}^{(k)} \delta(\mathbf{x} - \boldsymbol{X}_{t,s}^{(k)}). \tag{4.18}$$

It is important to mention that the algorithm just described is equivalent to the one presented by Bicacro et al. in [18]. However, a probabilistic interpretation of the final algorithm was presented in this section. To be more specific, the convolution of the image $f(\mathbf{x})$ that gathers all gazed positions with a Gaussian mask was also performed in the aforementioned work, but this step was only justified by the non-zero angle of maximum visual acuity.

4.4.3 Time-Dependent Eye Track Driven Selection

The Time-Independent ETDS procedure selects voxels according to a relevance criterion, which is the interest that each voxel raises to the physician. However, during diagnosis, the physician is able to use more information than the average intensities of different isolated brain regions. Therefore, the path taken by the physician's gaze point might also contain useful information, for example, by comparing the intensity levels in different brain regions. In order to simulate this behavior, a time-dependent probabilistic model of this path was constructed. The studied model is a first order model, since it considers only pairs of consecutive points to construct the probability mass function $P\left(\mathbf{x}_t, \mathbf{x}_{t+1}\right)$, from which one can randomly select pairs of voxels and then include into the feature set one or both voxel intensities – $V\left(\mathbf{x}_t\right)$ and $V\left(\mathbf{x}_{t+1}\right)$ – and/or transformations that highlight their difference – e.g. $\left(V\left(\mathbf{x}_t\right) - V\left(\mathbf{x}_{t+1}\right)\right)^2$. The algorithm studied in this thesis, TD-ETDS, used both $V\left(\mathbf{x}_t\right)$ and $\left(V\left(\mathbf{x}_t\right) - V\left(\mathbf{x}_{t+1}\right)\right)^2$ as features.

Before proceeding to further details, consider first a few changes in the notation in order to ease the presentation. Every pair of consecutive voxels analyzed by the physician will be available for the estimation of $P\left(\mathbf{x}_t, \mathbf{x}_{t+1}\right)$. The position of the first point will be denoted by $\boldsymbol{X}$ and the position of the consecutive one by $\boldsymbol{Y}$, and thus the pairs of voxels that form the dataset for this algorithm can be stated in the following way:

$$\left\{(\boldsymbol{X}, \boldsymbol{Y})_i\right\}^{(k)} \qquad \text{for} \qquad i \in \{1, \ldots, n_k\},\ k \in \{1, \ldots, K\}, \tag{4.19}$$

where n_k is the number of different pairs of consecutive voxels that can be extracted from (4.16) for patient k.

Since computing all entries of the probability mass function $P\left(\mathbf{x}, \mathbf{y}\right)$ is not a solution, due to memory limitations, its estimation and the extraction of the features used for learning were accomplished in two steps based on the conditional decomposition:

$$P\left(\mathbf{x}, \mathbf{y}\right) = P\left(\mathbf{x}\right) P\left(\mathbf{y}|\mathbf{x}\right). \tag{4.20}$$

First, the term $P\left(\mathbf{x}\right)$ was estimated using Parzen-Windows and using all gazed points $\boldsymbol{X}_i^{(k)}$, and then half of the desired number of voxels were sampled. Afterwards, for each sample $\tilde{\mathbf{x}}$ drawn, the second term $P\left(\mathbf{y}|\tilde{\mathbf{x}}\right)$ was estimated and the corresponding coupled voxel $\tilde{\mathbf{y}}$ subsequently extracted. Finally, for each pair of sampled positions $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ drawn, two features were added to the feature vector: $V(\tilde{\mathbf{x}})$ and $(V(\tilde{\mathbf{x}}) - V(\tilde{\mathbf{y}}))^2$.

In order to compute $\hat{P}\left(\mathbf{y}|\tilde{\mathbf{x}}\right)$, note that the Parzen-Window estimators for the marginal and joint distributions:

$$\hat{P}\left(\mathbf{x}, \mathbf{y}\right) = \frac{1}{\sum_k n_k} \sum_{i,k} \phi\left(\mathbf{x} - \boldsymbol{X}_i^{(k)}, h\right) \phi\left(\mathbf{y} - \mathbf{Y}_i^{(k)}, h\right) \tag{4.21}$$

$$\hat{P}\left(\mathbf{x}\right) = \frac{1}{\sum_k n_k} \sum_{i,k} \phi\left(\mathbf{x} - \boldsymbol{X}_i^{(k)}, h\right) \tag{4.22}$$

can be substituted into:

$$P\left(\mathbf{y}|\mathbf{x}\right) = \frac{P\left(\mathbf{x}, \mathbf{y}\right)}{P\left(\mathbf{x}\right)}, \tag{4.23}$$

yielding:

$$\hat{P}\left(\mathbf{y}|\tilde{\mathbf{x}}\right) = \frac{1}{\sum_{i,k} \phi\left(\tilde{\mathbf{x}} - \boldsymbol{X}_i^{(k)}, h\right)} \sum_{i,k} \phi\left(\tilde{\mathbf{x}} - \boldsymbol{X}_i^{(k)}, h\right) \phi\left(\mathbf{y} - \boldsymbol{Y}_i^{(k)}, h\right). \tag{4.24}$$

Note also that, in equation (4.21), the kernel for the 6-dimensional variable $\{\mathbf{x}, \mathbf{y}\}$ was factorized into two kernels. This can be done because the width parameter, h, is similar in every direction and does not introduce cross terms. Actually, it could be factorized into 6 kernels, one for each dimension.

Finally, a closer look at the expression (4.24) reveals that, similarly to what happened in the estimation of the time independent probability mass function, the estimation of the conditional distribution can be accomplished by the three-dimensional discrete convolution between the kernel $\phi\left(\mathbf{y}\right)$ restricted to integer coordinates, and a discrete function $f\left(\mathbf{y}\right)$ given by:

$$f\left(\mathbf{y}\right) = \frac{1}{\sum_{i,k} \phi\left(\tilde{\mathbf{x}} - \boldsymbol{X}_i^{(k)}, h\right)} \cdot \sum_{i,k} \phi\left(\tilde{\mathbf{x}} - \boldsymbol{X}_i^{(k)}, h\right) \delta\left(\mathbf{y} - \boldsymbol{Y}_i^{(k)}\right), \tag{4.25}$$

where $\delta\left(\mathbf{y} - \boldsymbol{Y}_i^{(k)}\right)$ represents a discrete unit impulse centered at the training sample $\boldsymbol{Y}_i^{(k)}$.

4.5 Conclusion

The chapter that ends here focused on the feature selection algorithms explored for dimensionality reduction purposes and, as argued in the introduction (Section 4.1), this stage is very important due to the high dimensionality of the input feature vectors and the comparatively small number of available training patterns characteristic of this specific problem, the CAD of AD. Five different techniques were presented. Three of them, PBCC, MIM and mRMR, are completely automated and rely directly on the dataset to estimate the value of the information that each feature holds, while the other two, TI-ETDS and TD-ETDS, are driven by medical expertise and are built using eye tracking data collected while an experienced physician was examining FDG-PET images, which makes these methods user dependent, or more specifically, dependent on the physician.

PBCC, MIM and mRMR presented several interesting connections between them. First, the structure of both PBCC and MIM algorithms is actually the same. They are both ranking methods, i.e., they both evaluate the usefulness of each isolated feature to the classification problem at hand. However, the usefulness criterion differs: on one side, PBCC uses the Pearson correlation coefficient which measures the linear dependence between each feature and the class label, while, on the other side, MIM uses mutual information that is able to account for any kind of relationship between variables. Nevertheless, both of these algorithms can suffer from the high redundancy between the intensity of nearby voxels, or of voxels located at symmetric regions of the brain. For instance, if one feature is considered useful, probably features associated with nearby voxels will also be considered useful, but the inclusion of both in the feature vector will probably not bring new information. The mRMR algorithm was used to circumvent this limitation and, as its name suggests, it mediates a trade-off between the relevance of a given isolated feature and its redundancy with the already selected ones. The main problem related to mRMR is its higher computational needs when compared to ranking algorithms.

On the other hand, the medically driven feature selection procedures, TI-ETDS and TD-ETDS, differ on one detail only. TD-ETDS takes into consideration the sequence through which the physician examined the FDG-PET scan, while TI-ETDS ignores this information. As a consequence, TD-ETDS is able to use as features, not only the intensities of voxels that caught most of the physician's attention, as in the TI-ETDS algorithm, but also the difference between intensities of consecutive voxels, which tries to mimic the physician when comparing different regions of the brain.

5 Classification and Performance Assessment

Contents

5.1 Introduction

The final step of any pattern recognition system is to learn a model from the training instances capable of correctly classifying future unseen data. As already mentioned, the high dimensionality of the feature vectors (even after the feature selection stage) suggests the use of a discriminative model. The Support Vector Machine (SVM) algorithm, perhaps, the most popular discriminative method for CAD both inside and outside of the AD research field, has already proven to achieve good generalization results even in almost empty spaces [58]. A comprehensive description of its concepts and the mathematics behind SVM will be given in section 5.2.

On a different topic, the assessment of any CAD system is very important for obvious reasons. In this work, the nested Cross-Validation (CV) technique was used to estimate, in an unbiased fashion, performance measures such as the classification accuracy, sensibility or specificity. Section 5.3 will present this method. Finally, section 5.4 concludes the chapter.

5.2 Support Vector Machines

5.2.1 Basic Concepts

Historically, Support Vector Machines represent a generalization to nonlinear models of the Generalized Portrait algorithm, introduced by Vapnik and Lerner [72] in 1963. However, only in the 90s, SVM took the form which is currently known, first in a paper authored by Boser, Guyon and Vapnik at the COLT '92 conference [73], and lastly introducing the notion of soft margin, vital for non-separable cases, in 1995 [74] (Cortes and Vapnik). Later, in 1998, Shawe-Tayler et al. [75] and Bartlett [76] presented the first rigorous bound to the generalization ability of the hard margin SVM, while for the soft margin version this bound was proposed by Shawe-Taylor et al. [77] in 2000.

The SVM algorithm excels for its simplicity. Consider first the binary classification of linearly separable data, as exemplified in Figure 5.1. This algorithm seeks the hyperplane that separates the data with maximum margin, i.e., the hyperplane that maximizes its distance to the closest training

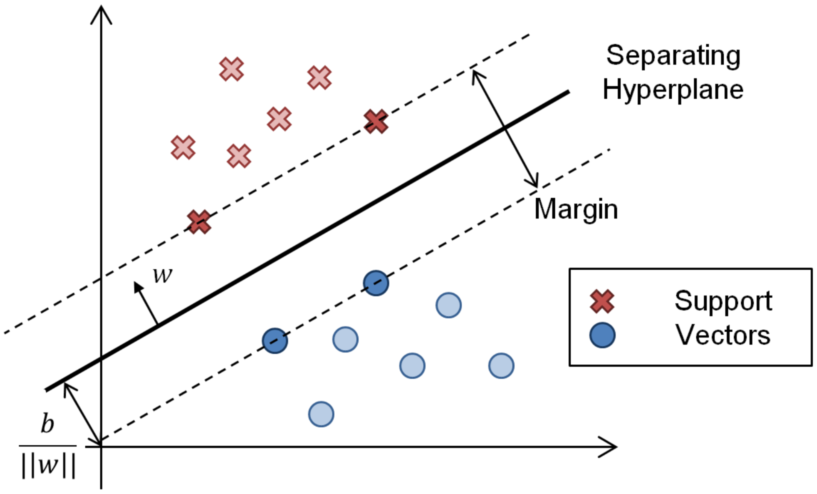


Figure 5.1: Illustrative example of the outcome of an SVM algorithm on a linearly separable binary problem. The optimal separating hyperplane maximizes the margin between the support vectors of each class.

vectors of each class, the so called support vectors.

On a different note, the good generalization ability that is observed in practice, in high dimensional spaces, was theoretically established by Vapnik et al. [73] for the first time. In this paper, the authors showed that the expected value of the probability of committing an error in a given unseen instance is bounded by:

$$E\left[p(e)\right] \leq \frac{E\left[\#\text{ Support Vectors}\right]}{\#\text{ Training Instances}}, \tag{5.1}$$

which does not explicitly depend on the dimensionality of the feature vector. Consequently, one might expect to achieve good performances even when the number of features is much higher than the number of training instances as long as the separating hyperplane can be constructed from a small number of support vectors, therefore alleviating the *curse of dimensionality*.

The simple *hard margin* concept lacks however expressive power and two extensions are often exploited. First, the use of *kernels* allows creating a nonlinear separation surface, by mapping the original *input space* into a typically higher dimensional *feature space*. Figure 5.2(a) illustrates a constructed example where the use of non-linear separation surfaces is crucial. Second, the use of the *soft margin* concept relaxes the separability constraint, allowing errors to be committed while trying to minimize them. This extension deals with situations where the data are not easily separable, unless an extremely complex kernel is applied, which is important because the use of such kernels typically results in an overfitted separation surface and, consequently, in poorer generalization abilities. An illustrative example is given in Figure 5.2(b). SVMs can also be modified to deal with more than two classes and to perform regression but those modifications are out of the scope of this thesis and will not be addressed.

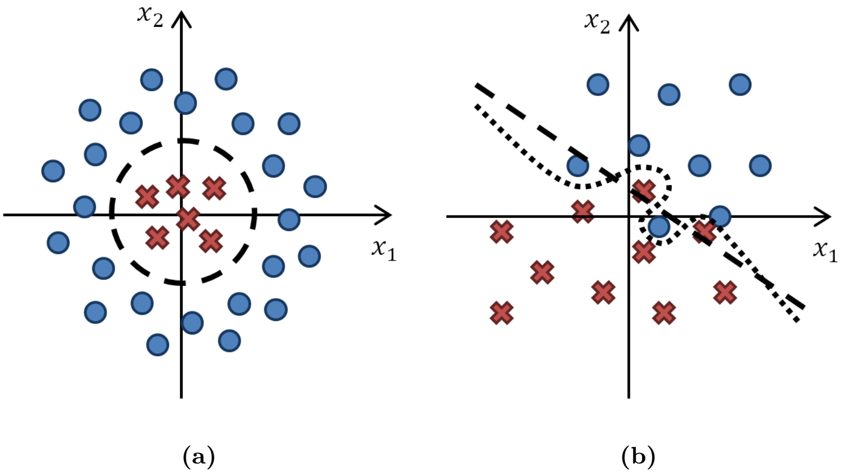


Figure 5.2: Illustrative examples of the need for SVM extensions. **(a)** No hyperplane is able to separate the data, despite the fact that a simple non-linear decision surface (a circumference) easily completes the task. This decision surface can be obtained by an SVM using the kernel obtained from the mapping $\phi(x_1, x_2) = x_1^2 + x_2^2$. **(b)** Non-separable problem. Two separation surfaces are illustrated. One is actually able to separate the classes but clearly overfits the data, while the other, a straight line, seems to be a better classifier, despite committing a few errors on the training instances.

5.2.2 Mathematics

The mathematics behind the implementation of the hard margin SVM [73,74] will now be covered. Consider a linearly separable problem where the training set $\mathcal{D}$ is formed by K instances, each one belonging to one of two classes:

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_K, y_K)\}, \quad \mathbf{x}_k \in \mathbb{R}^N, \quad y_k \in \{-1, 1\}, \tag{5.2}$$

where $\mathbf{x}_k$ is one instance of the N-dimensional feature vector built by one of the selection methods described in the previous chapter. It should be noticed that the iterator k for the training samples is here given in subscript, instead of in superscript as in chapter 4 for presentation purposes. A given hyperplane, which can be parameterized by its normal vector $\mathbf{w}$ and a constant b:

$$\mathbf{w} \cdot \mathbf{x} + b = 0, \tag{5.3}$$

is said to be a *separating hyperplane* if and only if the decision function given by:

$$f(\mathbf{x}) = sign\left(\mathbf{w} \cdot \mathbf{x} + b\right), \tag{5.4}$$

correctly classifies all training instances, i.e., if $f(\mathbf{x}_k) = y_k$ for all instances k, which can also be compactly rewritten in the following way:

$$y_k\left(\mathbf{w} \cdot \mathbf{x}_k + b\right) \geq 1 \quad \forall k. \tag{5.5}$$

Two details should be noticed in the previous inequality. First, the constant on the right-hand side could actually be any strictly positive number due to the fact that any hyperplane represented by $(\mathbf{w}, b)$ can also be represented by any scaled pair $(\lambda\mathbf{w}, \lambda b)$ with $\lambda \in \mathbb{R}^+$. The second and most important detail is that any separating hyperplane can be represented in such a way that equation (5.5) is met as equality for the nearest training sample(s) by changing the scaling factor λ.

Now, in order to select the best hyperplane from the infinite set of separating ones, its margin should be maximized. Bearing in mind that the distance between the hyperplane and the nearest vectors is given by:

$$\begin{aligned} d\left((\mathbf{w}, b), \mathbf{x}_k\right) &= \frac{y_k\left(\mathbf{w} \cdot \mathbf{x}_k + b\right)}{||\mathbf{w}||} && (5.6) \\ &= \frac{1}{||\mathbf{w}||}, && (5.7) \end{aligned}$$

one can conclude that the optimal hyperplane can be obtained by minimizing $||\mathbf{w}||$ or equivalently $\frac{1}{2}||\mathbf{w}||^2 = \frac{1}{2}\mathbf{w}^T\mathbf{w}$, under the constraints (5.5), i.e., by minimizing the following convex quadratic programming problem:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & y_k\left(\mathbf{w} \cdot \mathbf{x}_k + b\right) \geq 1 \quad \forall k \end{aligned} \tag{5.8}$$

The main method to solve this problem is based on its Lagrangian dual formulation which can be obtained as follows. First, consider the Lagrangian function associated with problem (5.8):

$$L(\mathbf{w}, b, \Lambda) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{k=1}^{K} \alpha_k \left[y_k\left(\mathbf{w}^T\mathbf{x}_k + b\right) - 1\right], \tag{5.9}$$

where $\Lambda = (\alpha_1, \dots, \alpha_K)$ is the vector of non-negative Lagrangian multipliers associated with constraints (5.5). Then, the infimum of $L(\mathbf{w}, b, \Lambda)$ with respect to $\mathbf{w}$ and b, quantity that the dual problem maximizes, can be determined using $\nabla_{\mathbf{w},b} L(\mathbf{w}, b, \Lambda) = 0$ and plugging the results into equation (5.9). Specifically, the conditions $\frac{\partial L(\mathbf{w},b,\Lambda)}{\partial \mathbf{w}} = 0$ and $\frac{\partial L(\mathbf{w},b,\Lambda)}{\partial b} = 0$ yield:

$$\mathbf{w} = \sum_{k=1}^{K} \alpha_k y_k \mathbf{x}_k \tag{5.10}$$

and

$$\sum_{k=1}^{K} \alpha_k y_k = 0, \tag{5.11}$$

respectively, and after some manipulation, the infimum of L can be given by:

$$\inf_{\mathbf{w},b} \{L(\mathbf{w}, b, \Lambda)\} = \sum_{k=1}^{K} \alpha_k - \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \alpha_k \alpha_l y_k y_l \left(\mathbf{x}_k^T \mathbf{x}_l\right). \tag{5.12}$$

The solution is then obtained by solving the dual maximization problem for the Lagrangian coefficients, which can be stated, using vector notation, in the following form:

$$\begin{aligned} \underset{\Lambda}{\text{maximize}} \quad & \Lambda^T \mathbf{1} - \frac{1}{2} \Lambda^T D \Lambda \\ \text{subject to} \quad & \Lambda \geq 0 \\ & \Lambda^T \mathbf{y} = 0 \end{aligned} \tag{5.13}$$

where $\mathbf{1} = (1, \dots, 1)$ is a K-dimensional unit vector, $\mathbf{y} = (y_1, \dots, y_K)$ is the vector of labels and D is a symmetric matrix with elements $D_{kl} = y_k y_l \left(\mathbf{x}_k^T \mathbf{x}_l\right)$ for k and $l \in \{1, \dots, K\}$. The dual problem is still a quadratic problem, but instead of scaling with the number of dimensions of the feature space, it scales with the number of training instances. In addition, one can notice that each input vector $\mathbf{x}_k$ always appears in a dot product with some other vector $\mathbf{x}_l$, a property whose usefulness will become clear later.

Finally, from optimization theory, more specifically from the complementary slackness condition of the Karush-Kuhn-Tucker theorem, one can conclude that, when the solution to the problem (5.13) is met, one of two possible situations holds true. If a given instance $\mathbf{x}_k$ is a support vector, then the associated Lagrangian multiplier α_k is non-negative, otherwise, α_k is zero. Consequently, the optimal hyperplane can be constructed as a linear combination of the support vectors using equation (5.10). In addition, the bias b can be found from the constraints (5.5) for the support vectors, since they are met as equalities for such training instances.

Following the historical evolution, the first extension of this concept was the introduction of *kernels*. As mentioned before, this extension is motivated by the fact that even if one dataset is not linearly separable, it can be separated by a nonlinear separation surface. Recall Figure 5.2(a). The introduction of *kernels* [73, 74] will solve this problem by using a mapping $\mathbf{z} = \phi(\mathbf{x})$ that transforms the original N-dimensional input space into a new N'-dimensional feature space, where an hyperplane will try to separate the new transformed data $\{(\phi(\mathbf{x}_k), y_k)\}$.

Two complications can be identified at this stage. The first is computational and is associated with the dimension of the feature space, which can even be infinite for some types of kernels, precluding a naive straightforward approach. However, after replacing all occurrences of $\mathbf{x}$ by $\phi(\mathbf{x})$, one can see that each time a given $\phi(\mathbf{x}_k)$ appears, is in a dot product with another $\phi(\mathbf{x}_l)$. As a consequence, one only needs to define the inner product in the feature space, without having to explicitly compute the mapping of the input vectors. Specifically, the elements of matrix D of the dual problem (5.13) become $D_{kl} = y_k y_l \left(\phi(\mathbf{x}_k) \cdot \phi(\mathbf{x}_l)\right)$ and the decision function $f(\mathbf{x})$ can also be expressed in terms of inner products, just by writing the vector $\mathbf{w}$ as the linear combination of the support vectors in the feature space, $\mathbf{w} = \sum_{k=1}^{K} \alpha_k y_k \phi(\mathbf{x}_k)$, and then replacing $\mathbf{w}$ into equation (5.4), yielding:

$$f(\mathbf{x}) = sign\left(\sum_{k=1}^{K} \alpha_k y_k \left(\phi(\mathbf{x}_k) \cdot \phi(\mathbf{x})\right) + b\right). \tag{5.14}$$

The inner products are given by the so called kernel function:

$$K(\mathbf{x}_k, \mathbf{x}_l) = \phi(\mathbf{x}_k) \cdot \phi(\mathbf{x}_l). \tag{5.15}$$

The second problem arises with the choice of the kernel. In low-dimensional spaces, one might be able to imagine a mapping function that would separate the training set in the feature space, but this task becomes impossible when the dimension of the input space starts to grow. Fortunately, several kinds of kernels with good generalization ability were already identified. The RBF kernel, defined in equation (5.16), is one of them and it has been extensively used for the CAD of AD [16,18,19,30,33].

$$K(\mathbf{x}_k, \mathbf{x}_l) = \exp\left\{-\gamma \left|\left|\mathbf{x}_k - \mathbf{x}_l\right|\right|^2\right\} \tag{5.16}$$

So far, the training data was assumed to be separable either on the input space or in the feature space. However, if the data cannot be fully separated without committing a small number of errors, the dual problem becomes unbounded and no solution can be found. In addition, the use of complex kernels to separate the data often leads to poorer classification performance. Recal Figure 5.2(b). To handle this problem Cortes and Vapnik [74] relaxed the constraints (5.5), introducing a positive slack variable ξ_k in each one:

$$y_k \left(\mathbf{w} \cdot \mathbf{x}_k + b\right) \geq 1 - \xi_k \quad \forall k. \tag{5.17}$$

Those slack variables, which quantify the errors committed by each vector, were then weighted in the cost function in order to keep them as small as possible. The new primal problem can therefore be stated as follows:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{k=1}^{K}\xi_k \\ \text{subject to} \quad & y_k \left(\mathbf{w} \cdot \mathbf{x}_k + b\right) \geq 1 - \xi_k \quad \forall k \\ & \xi_k \geq 0 \qquad\qquad\qquad\quad\; \forall k \end{aligned} \tag{5.18}$$

where C is a tuning parameter that controls the cost of misclassification. This optimization problem, which is still convex and quadratic as the original one, is often solved by exploiting its dual representation, which can be obtained following a reasoning similar to the separable case. At the end, the dual problem takes the form:

$$\begin{aligned} \underset{\Lambda}{\text{maximize}} \quad & \Lambda^T\mathbf{1} - \frac{1}{2}\Lambda^T D \Lambda \\ \text{subject to} \quad & 0 \leq \Lambda \leq C \\ & \Lambda^T\mathbf{y} = 0 \end{aligned} \tag{5.19}$$

In this work, the dual problem of the soft margin SVM algorithm was solved numerically using LIBSVM, a publicly available software developed by Chang and Lin [78] and available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

5.3 Nested Cross-Validation

In supervised learning, in which the main objective is to classify future unseen data, one is usually interested in evaluating the performance of the proposed classifiers. However, if this assessment was performed using the same instances that were used to train the classifier, it is known that an optimistically biased estimate would be obtained. In addition, one frequently faces the task of choosing certain parameters, such as the RBF parameter γ or the parameter associated with the SVM soft margin C, that may influence the final classification accuracy. In order to select the best model and to subsequently evaluate it, two different situations often arise. The first, when there are sufficient data, the input vectors can be grouped into three disjoint sets, namely, the training, test and validation set. The training set is used to learn the model, while parameters are selected according to the accuracy obtained in the validation set. Then, after all parameters have been selected and the final classifier constructed, its generalization ability can be evaluated using the test set, which was never used in the learning stage, therefore guaranteeing an unbiased estimate. The second situation occurs when the number of available input patterns is small and it is not advisable to leave data out from the training stage since the feature space is already sparsely occupied. One method that circumvents this difficulty is the cross validation (CV) procedure. In k-fold CV, the dataset is first randomly partitioned into k disjoint sets. In the current study, since all classes had exactly the same number of instances, the same proportion (50/50) was forced to each partition. Next, one of the k sets is chosen to be the validation set, while the others are used to train the classifier. This procedure is repeated k times, with all partitions being selected exactly once as the validation set. Measures such as the classification accuracy, sensitivity or specificity can therefore be estimated by comparing the classification of each training vector, computed when it was part of the validation set, with the true labels.

Now, in order to tune the model parameters, one might be tempted to repeat the CV procedure several times, one for each parameter setting, select the setting that achieved highest performance, and report the corresponding accuracy as the measure of the classifier generalization ability. However, it was shown by Varma and Simon [79] that the error of the estimate resultant for this procedure is significantly biased because the patterns that were used to evaluate the classifier were also used to learn the final model, specifically to perform parameter selection. In the same paper, Varma and Simon proposed the so called nested CV procedure which solves this limitation. This method partitions the initial data into k' disjoint sets. Then, in each iteration, one set is left out as the test set, while the others enter in several CV procedures, one for each setting, from which the best setting is chosen. Since the aim of the inner CV is to search for the best model, the same partitioning was used in all CV procedures, guaranteeing that results attained for the different settings were comparable. Finally, a model is learned with the chosen parameters and using all vectors except the ones in the test set, which are used to evaluate the aforementioned model. This iteration is repeated k' times using a different

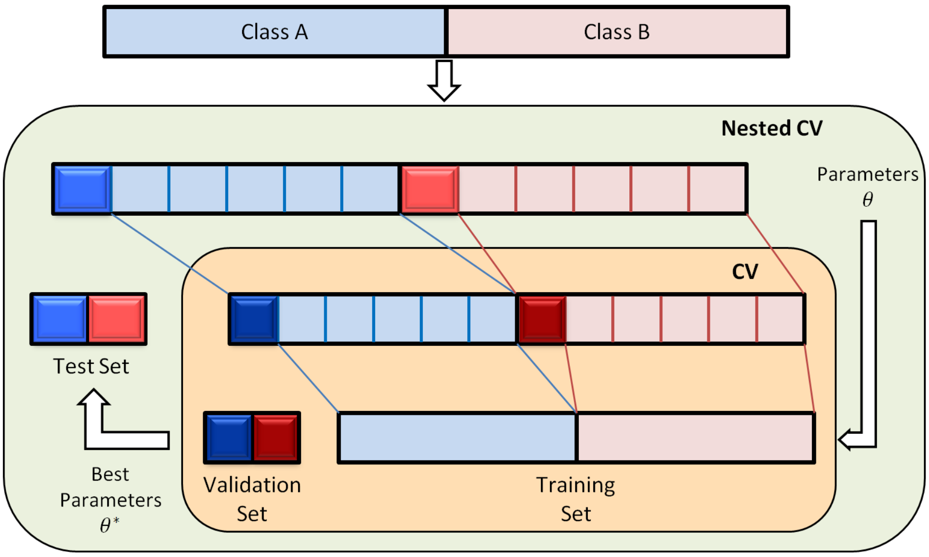


Figure 5.3: Partitioning performed in the nested cross-validation procedure. The initial database belongs to two classes, class A (blue) and class B (red). Within the nested CV procedure, the dataset is partitioned into k' folds. A different fold is used as the test set, in each iteration, and the remaining ones enter several CV procedures. Within each CV, the dataset is partitioned into k folds and, in each iteration, one is used as the validation set, while the others constitute the training set. Parameter selection is performed based on the classifications attained in the validation sets, while performance assessment is performed based on the classification attained in the test sets.

partition as the test set at a time and, at the end, the classification of each instance, obtained when it was part of the test set, is used to assess the system's performance. Figure 5.3 schematizes the partitioning done in the nested CV procedure.

5.4 Conclusion

In this chapter, the learning algorithm used to perform the automated diagnosis of Alzheimer's disease was covered in detail, namely, the Support Vector Machine algorithm which is known to deal well with high dimensional problems i.e., to be robust to the *curse of dimensionality*. It is once again stressed that this property is extremely important to the classification task at hand, since for PET data, the number of input patterns is much smaller than their dimensionality. In addition, two commonly used extensions to the basic hard margin SVM algorithm were covered, the RBF kernel and soft margins.

On a different topic, model selection and performance assessment of proposed classifiers (composed by one feature extraction algorithm, one feature selection technique and the SVM with or without the RBF kernel) were carried out using the nested CV procedure which was also described in this chapter. It is important to note that, in each iteration, the input data to both feature selection and learning stages comprise only the instances in the training set of the inner CV. This procedure guarantees almost unbiased estimates of performance accuracy, sensitivity and sensibility. In fact, those estimates are known to be slightly pessimistic because not all patterns were used to derive the model that was applied to the test sets.

6 Experimental Results

Contents

6.1 Introduction

Chapters 3 to 5 focused on the three most important building blocks of a successful CAD system: feature extraction, feature selection and learning. The results associated with the CAD system will now be presented, starting by the description of the neuroimaging data used to learn and test the classifiers in section 6.2, followed by a brief practical discussion of each feature extraction and feature selection algorithm in section 6.3, which will justify important implementation choices summarized in section 6.4. Finally, section 6.5 will present the performance of the implemented classifiers.

6.2 Dataset – Neuroimaging Data

Neuroimaging data used to train and test all proposed classifiers were taken from the ADNI database. ADNI is a multisite study committed to determine the sequence of events that take place in the progression of AD, as well as to establish standards for imaging and biomarker collection and analysis [80]. This study started in 2004 and recruited, during its first grant ADNI1, 819 participants, from which 229 were elderly control subjects, 402 suffered from MCI and the remaining 182 from AD. At the baseline visit, about half of the subjects were submitted to an FDG-PET scan, as well as at the follow-up visits that occurred at months 6, 12, 18, 24, 30 and 36, except for AD patients who stopped being followed after the second year [81].

In the present study, all CN, MCI and AD subjects whose FDG-PET scans were available were considered, as long as each person's CDR score met the following restrictions: 0 for normal controls, 0.5 for MCI patients and 0.5 or higher for AD patients, resulting in an intermediate dataset composed by 70, 104 and 59 subjects, respectively. The dataset herein utilized was then built by selecting randomly 59 patients from each group (except for the AD group where all subjects were retained). The reduction of the number of instances was conducted in order to reduce the number of PET scans to be examined by the physician. Table 6.1 summarizes important clinical and demographic information for each group, such as age, sex and MMSE scores.

Table 6.1: Description of the studied groups. Values are presented in "Mean ± Standard Deviation" format.

Group	AD	MCI	CN
Number of patients	59	59	59
Age	78.3 ± 6.6	77.7 ± 6.9	77.4 ± 6.6
Sex (% of Males)	57.6	67.8	64.4
MMSE	19.6 ± 5.1	25.8 ± 3.0	29.2 ± 0.9

Image acquisition was conducted using one of the following protocols [82]:

- Dynamic - six five-min frames, performed 30 to 60 minutes after FDG injection.
- Static - a single 30-min frame, 30 to 60 minutes post-injection.
- Quantitative - a 60 minutes dynamic protocol consisting of 33 frames, starting at the moment of tracer injection.

Since neuroimaging data available are produced by different PET scanner models, ADNI's investigators made an effort to standardize the resulting images in what regards resolution, orientation, dimensions, file format, etc.. The following consecutive preprocessing steps were performed:

1. Dynamic co-registration: Dynamic and quantitative protocols create a plurality of frames which may not be align due to distortions such as patients' motion. Therefore, all frames were co-registered to the first one (the base frame), ensuring a uniform representation of all moments of acquisition. The resulting images kept the base frame size and voxel dimensions and remained with the original spatial orientation, which was called "native" space.

2. Averaging: A single 30-min frame was generated by averaging the 6 five-min frames for the dynamic protocol or the last 6 frames for the quantitative one, using the image set obtained in the previous step. Images were still in the "native" space.

3. Image and voxel size standardization: Each co-registered, averaged image was reoriented and mapped into a standard 160×160×96 voxel grid, having 1.5 mm cubic voxels, where the orientation was chosen to align the anterior-posterior axis and the AC-PC line.

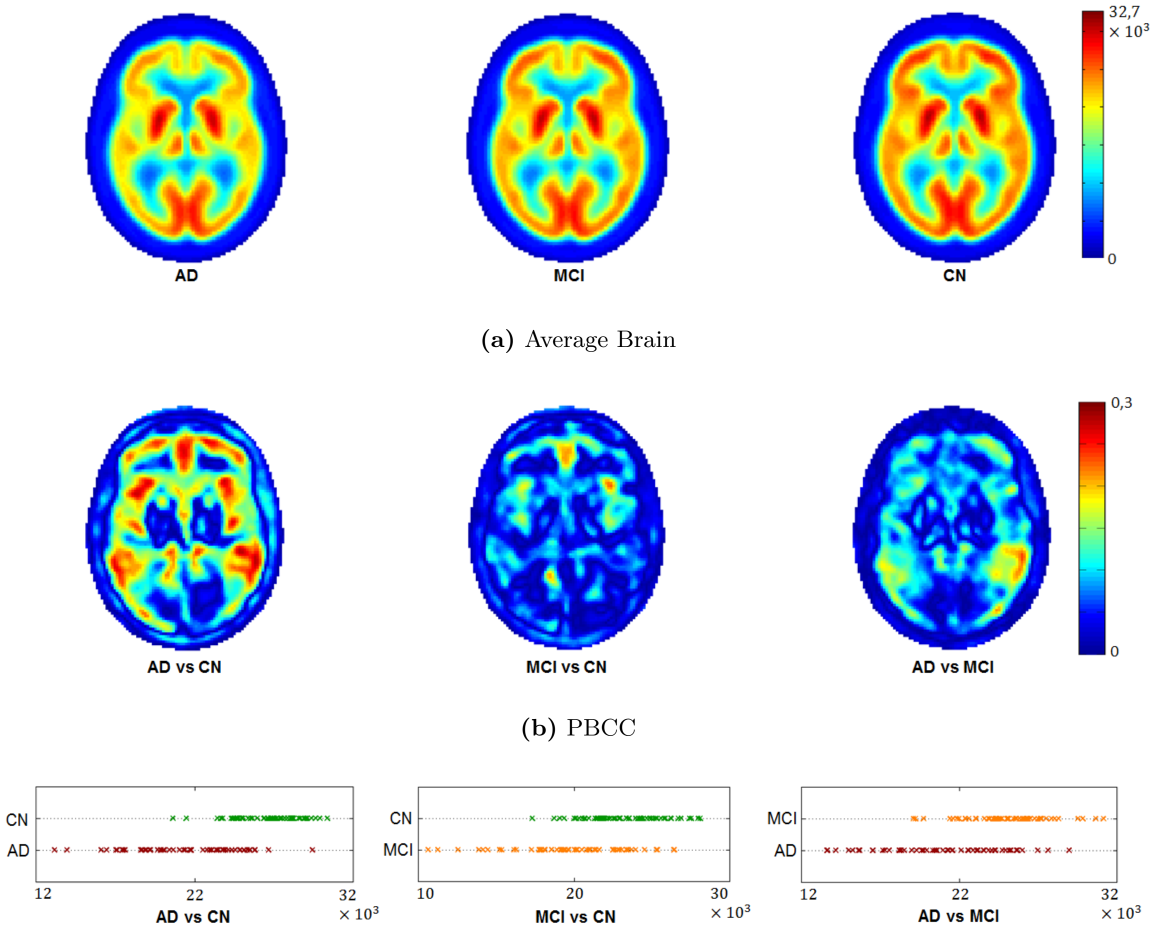
4. Resolution standardization: After the first three steps, each image set was smoothed in order to obtain a uniform isotropic resolution of 8 mm FWHM, using scanner-specific filter functions.

5. Talairach warp: PET volumes were also registered to the Talairach space, so that equal positions in the voxel grid match the same anatomical positions in all subjects, regardless of individual differences in the brain size or its overall shape. A 128×128×60 voxel grid was generated from this non-linear warping

6. Intensity normalization: Finally, all images were individually normalized, resulting in a complete span of the [0 32700] interval.

6.3 Practical Discussion on Feature Extraction and Selection Algorithms

FDG-PET is the central biomarker used in this study. As it was already underlined, VI features extracted directly from FDG-PET scans measure the FDG uptake in each brain location and since Alzheimer's disease causes a reduction of brain activity, then VI should incorporate valuable information, useful for automatic classification. In order to visualize the usefulness of this type of feature, Figure 6.1(a) shows an axial cut of the average brain volume computed for the three classes. A closer look easily reveals brain regions where the intensity is lower for AD patients and higher for CN subjects. To quantify the separation ability of each isolated VI feature, several criteria have been proposed, some of them were already described in this thesis, namely PBCC and MI. Figure 6.1(b) shows the same slice presented in Figure 6.1(a) but now each voxel holds the PBCC value, i.e., its own correlation with the class label. Finally, Figure 6.1(c) shows the feature that achieved the highest PBCC value in each classification task. From the last two rows of Figure 6.1, one can safely foresee that automatic diagnostic involving MCI subjects will be more difficult than the AD vs. CN problem. The literature review



(a) Average Brain

(b) PBCC

(c) Best VI Feature

Figure 6.1: Separability of VI features. **(a)** The slice $s = 30$ of the average brain computed for the AD, MCI and CN groups is illustrated. **(b)** The same slice is presented, but each voxel now contains its own correlation coefficient (PBCC) with the class label. **(c)** Intensities of the voxel with the highest PBCC score for all patients in the dataset.

conducted in chapter 2 already showed this fact, where the performance of recognition systems applied to the diagnosis of AD (against normal controls) was significantly better. Nevertheless, it should be noticed that even the best feature in the AD vs. CN task does not have an high correlation with the class label and, therefore, in order to efficiently separate the data, it will probably be necessary a large number of features working together and complementing each others.

Regarding the PET images' scale-space, the pyramid was expanded up to level four, with a subsampling factor of two in each direction. The use of more layers was not considered since the number of voxels in the fourth level is already very small, and, therefore, it is expected that this layer (and obviously the subsequent ones) will lead to poorer separation ability. In addition, one brain mask was constructed for each level using a similar approach to the one described in section 3.2. The total number of intracranial features retained in each layer is presented in Table 6.2.

On a different topic, two parameters had to be tuned for the local variance feature: the number of neighbors P and radius R of the sphere where the neighbors are located. The most important

Table 6.2: Number of VI features in each layer of the scale-space of the brain images.

Layer (l)	Whole Volume Dimensions	Total Number of Features	Number of Intracranial Features
0	128 × 128 × 60	986 040	319 441
1	64 × 64 × 30	112 880	46 001
2	32 × 32 × 15	15 360	7 348
3	16 × 16 × 8	2 048	1 404
4	8 × 8 × 4	256	252

parameter is the radius because it controls the scale at which local contrast is being measured. The number of neighbors should not have a major impact as long as it is set high enough. A reasonable rule of thumb is to choose for P the number of voxels that lie on the surface of a cube of size $2R + 1$.

The same parameters, R and P, are also associated with Local Binary Patterns, but, in this situation, two restrictions arise. One is related to the exponential time complexity of labeling all possible patterns, which was performed offline, analyzing all of them, one by one. This restriction is more problematic for the 3D case, since, in two dimensions, one can easily enumerate all uniform patterns without having to examine non-uniform ones, while for the 3D case that is not possible. The second constraint is associated with the number of uniform and rotation invariant LBPs which can not be too high (at most a few hundreds) for the sake of histogram stability, as it was stressed in sections 3.5.1 and 3.5.2. Tables 6.3 and 6.4 present some statistics about 2D and 3D LBPs, respectively, and their application to the database herein utilized.

The analysis of Tables 6.3 and 6.4 reveals important information about the influence of both parameters, which will be useful to specify a tuning strategy. First, one can see that the reduction in the number of patterns achieved by the uniformity concept is extremely large. Nevertheless, those patterns are characteristic of the brain volumes since they arise with higher frequency than non-uniform patterns. For instance, in the 2D case and for the setting $(P, R) = (24, 3)$, uniform LBPs represent only 0.0033% of the total number of possible patterns (554 out of 17 million) and, yet, they account for 84.6% of the LBP instances found in the database. In the 3D case, the results are slightly worst since, not only the number of uniform patterns increases, but the percentage of occurrence decreases. Nevertheless, the numbers are still impressive, with 0.035% of the patterns occurring with a frequency of 28.2% in the worst scenario presented. In general, one can verify that the percentage of occurrence typically decreases with P and N. On the other hand, the number of uniform and rotation invariant patterns never reaches prohibitive values for the settings considered in Tables 6.3 and 6.4. The worst result arises in the three-dimensional case with $P = 24$, and can be explained by the use of an approximation to the equidistant neighbor positioning on the sphere. Also in the 3D case, the parameter P was not studied for more than 24 samples because the number of possible patterns would make it virtually impossible.

That being said, one needs to define a good strategy in order to efficiently tune these parameters. First, since LVAR presents no limitations, R should be tested for different values to search for patterns

Table 6.3: Statistics of 2D LBPs. Only LBPs built at intracranial positions were considered when computing the percentage of occurrence of uniform patterns. Abbreviations: U-LBPs – Uniform LBPs; U-RI-LBPs – Uniform and rotation invariant LBPs;

P	Number of LBPs (2^P)	Number of U-LBPs	Percentage of Occurrence of U-LBPs (R=1,R=2,R=3,R=4,R=5,R=6)	Number of U-RI-LBPs
8	256	59	(97.9, 93.8, 88.7, 85.4, 83.7, 82.4)%	10
16	66×10^3	242	(97.7, 92.1, 89.2, 79.7, 74.5, 71.5)%	18
24	17×10^6	554	(97.6, 91.0, 84.6, 77.1, 73.3, 68.3)%	26
32	4×10^9	1 262	(97.6, 91.5, 84.3, 77.9, 71.5, 67.5)%	34
36	69×10^9	1 262	(97.6, 91.3, 84.0, 77.7, 71.6, 66.9)%	38
48	281×10^{12}	2 258	(97.6, 90.9, 84.0, 76.9, 71.3, 66.8)%	50

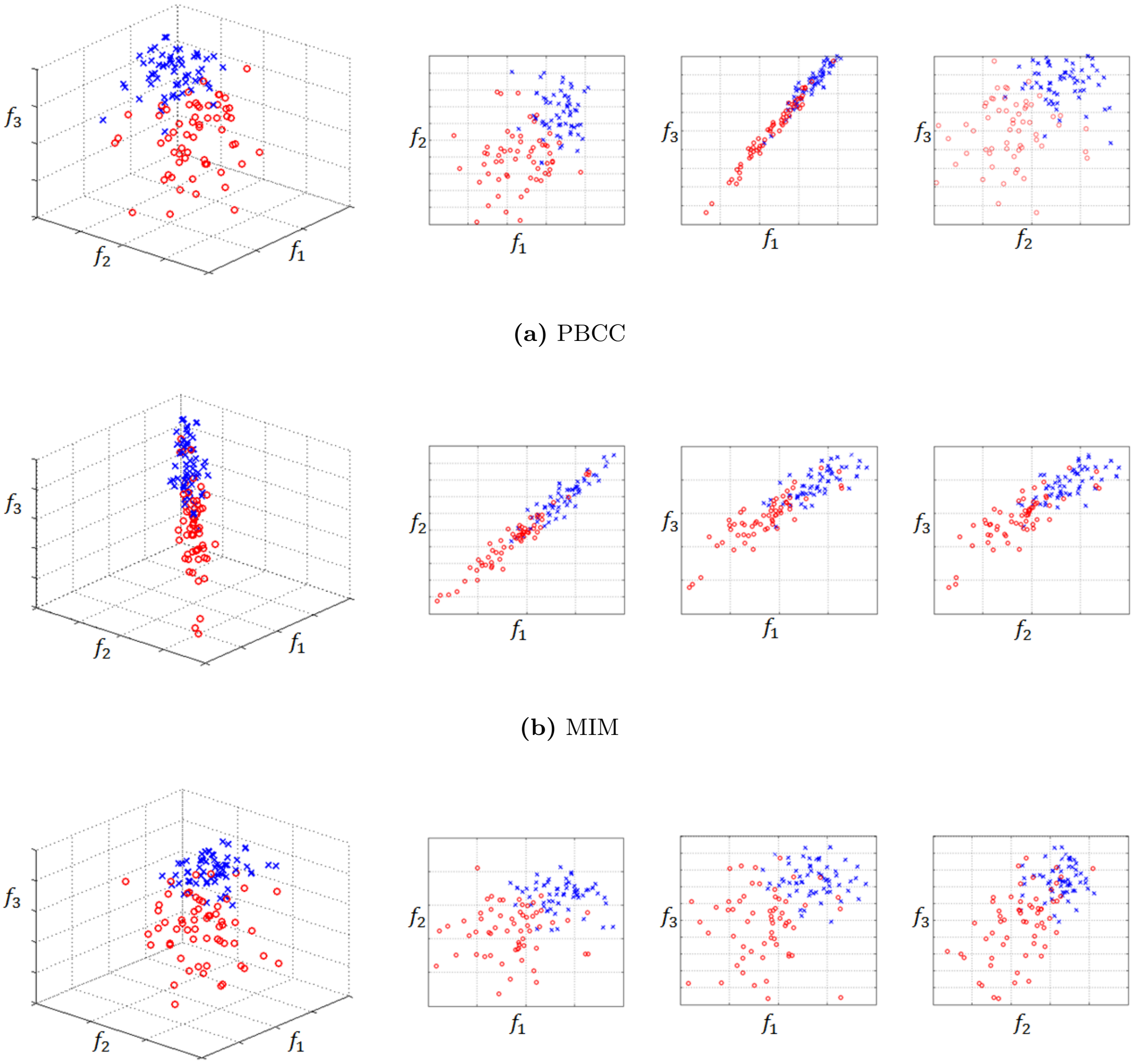
Table 6.4: Statistics of 3D LBPs. Only LBPs built at intracranial positions were considered when computing the percentage of occurrence of uniform patterns. Abbreviations: U-LBPs – Uniform LBPs; U-RI-LBPs – Uniform and rotation invariant LBPs;

P	Number of LBPs (2^P)	Number of U-LBPs	Percentage of Occurrence of U-LBPs (R=1,R=2,R=3,R=4,R=5,R=6)	Number of U-RI-LBPs
8	256	99	(92.4, 89.1, 79.3, 63.9, 63.6, 63.5)%	12
12	4×10^3	594	(86.7, 72.4, 69.3, 57.1, 55.4, 53.8)%	27
24	17×10^6	5 949	(62.3, 53.6, 42.3, 33.8, 30.4, 28.2)%	211
36	69×10^9	–	–	–

at different scales, while for each radius a single value for P should be used, computed using the rule of thumb presented before, and therefore reducing the number of settings to test. The same reasoning applies to the two-dimensional LBPs, but since a 2D neighborhood is being used, in this situation, the number of neighbors should be chosen according to the number of pixels located in the perimeter of a square (instead of a cube) of size $2R + 1$. At last, parameter P is limited to at most 24 samples for the 3D case, and therefore the previous rule of thumb can not be applied, meaning that a sparser neighborhood must be used. Since the neighbors are already sparsely distributed over the sphere, P was chosen to always be as high as possible, i.e., 24 for all tested radii.

As regards feature selection algorithms, only one parameter common to all methods had to be tuned – the number of features N to select. This tuning procedure was conducted using an exponential grid search approach. On a different note, in order to visualize the selection criterion used by the three fully automated selection procedures, the best three features selected by each algorithm (PBCC, MIM, mRMR) to differentiate subjects from the AD and CN groups are illustrated in Figure 6.2. Each point represents one subject.

Bearing in mind that the number of voxels to select will be (much) higher than three, one can confirm from Figure 6.2 that mRMR is superior to PBCC and MIM, as expected. On the first three features, mRMR did not choose a single pair of highly correlated features, as opposed to the other techniques. For instance, the best and the third best features selected by PBCC are actually

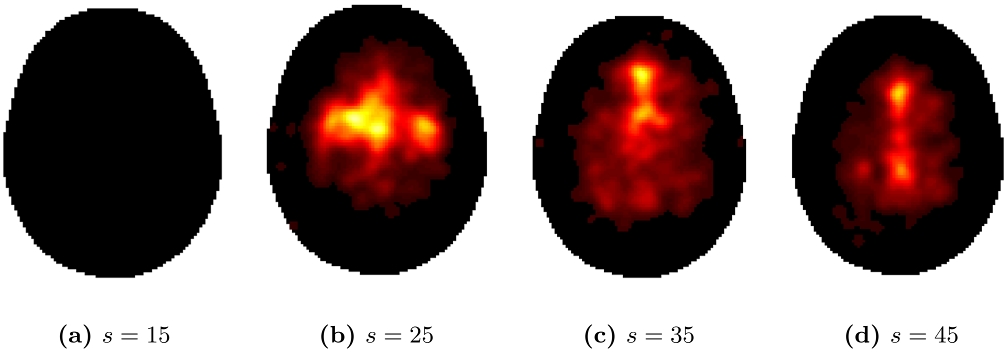


(a) PBCC

(b) MIM

(c) mRMR

Figure 6.2: Best three features (f_i is the ith best) selected for the AD (red circles) vs. CN (blue crosses) classification task by: **(a)** PBCC; **(b)** MIM; **(c)** mRMR.



(a) $s = 15$ **(b)** $s = 25$ **(c)** $s = 35$ **(d)** $s = 45$

Figure 6.3: Probability function used to draw features for the TI-ETDS procedure. Each entry of the aforementioned probability function is associated with one brain position. Several axial cuts are depicted.

neighboring voxels.

Two medically informed techniques were also implemented. The first one, TI-ETDS, draws brain voxels from a probability function that is associated with the amount of time spent by the physician analyzing each region of the brain. Several slices, where each voxel represents its own probability of being selected by this method for the AD vs. CN task, are shown in Figure 6.3. The comparison between Figures 6.3 and 6.1(b) reveals, surprisingly, that PBCC and TI-ETDS will preferably choose features form different regions on the brain.

As regards the TD-ETDS technique, the probability mass function $P(\mathbf{x}_t, \mathbf{x}_{t+1})$ was partially constructed using the decomposition described in section 4.4.3. Notice that, since the physician is restricted to watch one axial cut at a time, the z coordinate at times t and $t+1$ never changes. Figure 6.4 shows one slice of the total probability mass function. From this probability function, a feature vector was constructed by randomly choosing $N/2$ pairs of consecutive voxels, and for each pair adding the two following features: $V(\mathbf{x}_t)$ and $(V(\mathbf{x}_t) - V(\mathbf{x}_{t+1}))^2$.

The window width parameter of the Parzen-Window estimator, h, was set to 1.5 voxels in all cases which means that each sample influences significantly the selection likelihood of voxels located at distances smaller 3. This parameter was not tuned in order not to increase the computational burn, and mainly because its influence on the performance of the final classifier should be small, provided that h is neither too small nor too large.

Finally, the SVM kernel parameters were also tuned using an exponential grid search approach. In order to define the search range for each parameter, the system was first evaluated for a wide number of settings, from 2^{-18} to 2^{18} for both C and γ parameters, for all types of feature and selection procedures using different values of N. The assessment was carried through several cross validation procedures, one for each setting. Figure 6.5 presents, as an example, the classification accuracy attained for the AD vs. MCI problem, using PBCC to select 50 VI features and varying the kernel parameters. It is important to note that if, for instance, the range of the grid used to tune C in the linear case (Figure 6.5(a)) did not include values around 2^{-6}, the system's performance would be harmed. A similar concern arises with the RBF kernel (see Figure 6.5(b)).

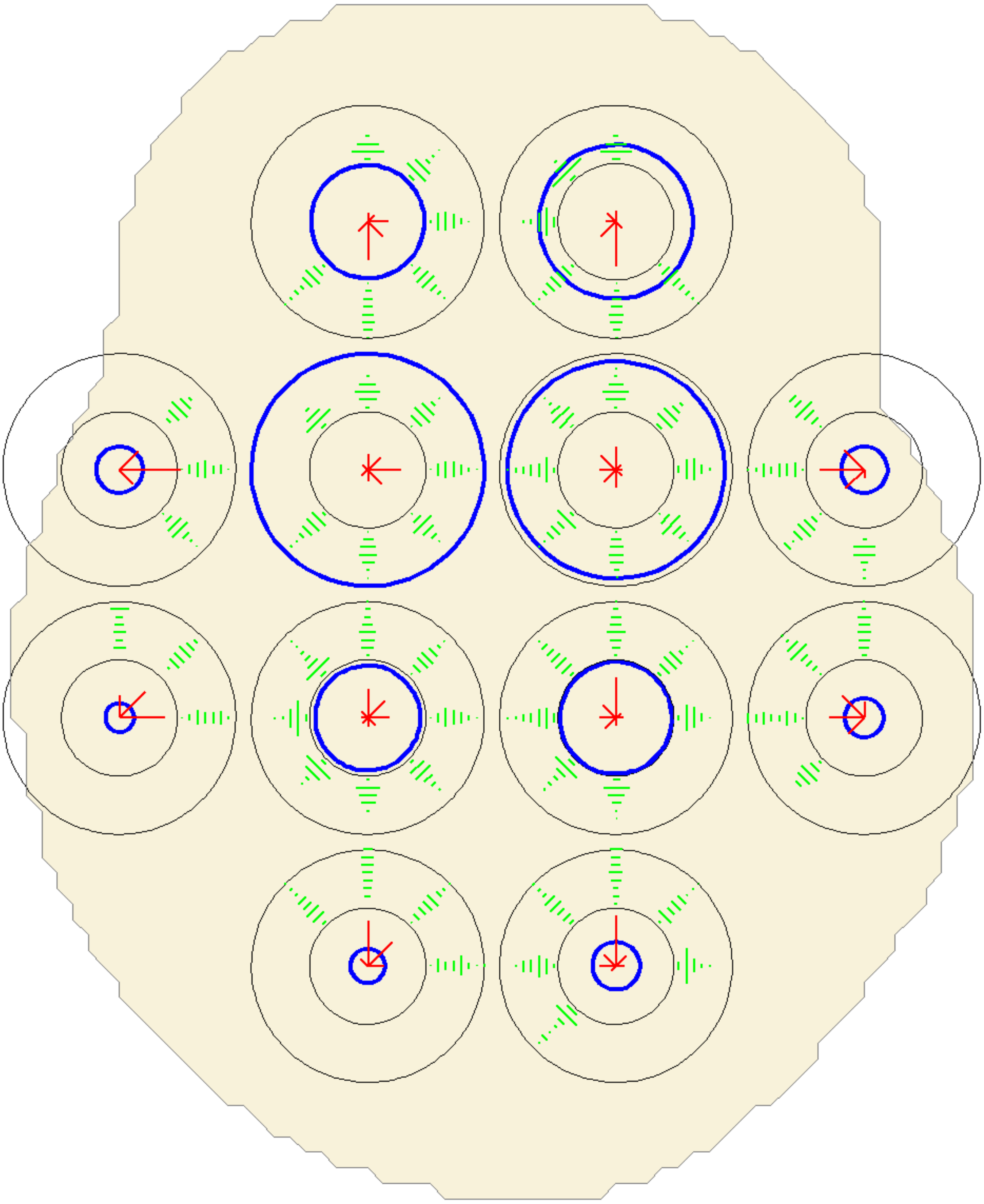
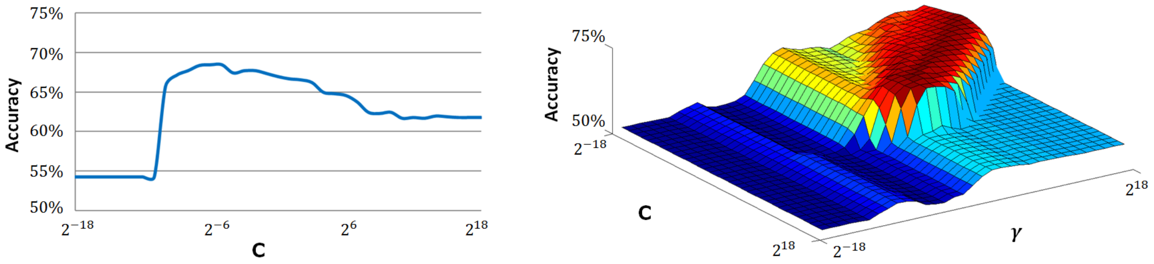


Figure 6.4: Representation of the probability function $P(\mathbf{x}_t, \mathbf{x}_{t+1})$ used to draw features for the TD-ETDS procedure. Each structure is associated with the brain position located at the center of the structure. The radius of each blue circle is proportional to the probability $P(x_t, y_t)$. The length of red arrows is proportional to $P(\theta|x_t, y_t)$ where θ is the direction of gaze from point (x_t, y_t) to point (x_{t+1}, y_{t+1}). The length of each green bar is proportional to $P(d|\theta, x_t, y_t)$ where d is the distance between the aforementioned points. The value of d increases from the inner black circumference to the outer one. Note that only a few structures are illustrated in this image, but the probability function can actually be computed for any two points $\mathbf{x}_t, \mathbf{x}_{t+1}$. Only one axial cut ($s = 25$) is depicted.



(a) Linear Kernel

(b) RBF Kernel

Figure 6.5: Classification accuracy for the AD vs. MCI problem, varying the kernel parameters. Results were produced using PBCC to select 50 VI features.

6.4 Experimental Design

The present work had two main objectives: to study different feature selection algorithms and to evaluate alternative features of the FDG-PET image.

The comparison of all feature selection algorithms was conducted using the highest resolution layer of the PET image's scale-space and a linear kernel for the SVM algorithm. The RBF kernel was not employed here because a study carried by Bicacro et al. [18] showed that this particular kernel does not improve classification accuracy when VI features are used. Moreover, an additional dummy algorithm, which performs the selection in a completely arbitrary manner, was implemented for comparison purposes. It will allow us to assess if the selection algorithm is boosting the system's performance by choosing the best features first or if those results only reflect the "average" separation power of the type of feature in use.

On the other hand, all three types of feature, VI, LVAR and LBPs, were evaluated. To be able to extract and compare the best possible performances, both linear and RBF kernels were tested, as well as the three fully automated feature selection procedures, PBCC, MIM and mRMR. Medically driven procedures were not considered in this experiment because they could not be directly used with Local Binary Patterns. Remember that features produced by LBPs represent frequencies of occurrence of different LBP labels in different regions of the brain, while both ETDS procedures are only able to select brain positions. As regards VI features, all five layers of the scale-space were assessed independently, so that one could conclude which is the layer of minimum resolution that achieves better or, at least, comparable performances. Also, both two-dimensional and three-dimensional LBPs were tested in order to assess the performance gain obtained by the 3D upgrade. Now, if all possible combinations of feature extraction, feature selection and SVM kernel were considered for the three classification tasks between the AD, MCI and CN groups, one would need to evaluate 144 CAD systems. A simpler approach was undertaken. First, a default feature selection algorithm and SVM kernel were set constant, while all levels of the scale-space were being tested. At the end, only one was chosen to proceed to the next phase. Then, the best selection algorithm was sought for each type of feature, in each classification problem, still with the default SVM kernel fixed. Finally, both kernels were tested for the feature extraction and selection procedures chosen by the previous steps.

MIM was chosen as the default selection algorithm for two main reasons: it is computationally efficient because it is a ranking algorithm, and it was preferred over the technique based on correlation coefficients due to its theoretical advantage, i.e., mutual information accounts for any kind of dependency between variables, while PBCC only quantifies linear relationships.

The linear kernel was set as the default kernel, most importantly, due to its lower computational needs. Actually, although the time spent in the training stage by the SVM algorithm with one predefined combination of kernel parameters is similar to both linear and RBF kernels, the tuning of fewer parameters makes the former more attractive. Note this optimization is performed using a grid search approach within the nested CV procedure and the linear kernel has one parameter less to optimize.

It should be stressed that beyond the kernel parameters, C and γ, the number of selected features, N, was also tuned within the nested CV, unless stated otherwise. In addition, the assessment of each CAD system was always estimated as the average of 10 nested CV runs, in order to diminish statistical variations. Other algorithm specific parameter specifications are described below:

- **Feature Selection Algorithms** – Classification was performed with $N \in \{50, 100, 250, 500, 1000, 2500, 5000, 10000, 25000, 50000\}$, except when mRMR was involved. For this method, the same progression was used, but the maximum value was chosen (differently for each type of feature) so that each nested CV run could be completed in less than a day. More specifically, N was allowed to assume values up to 500 for the VI and 2D-LBP types of feature and up to 100 for LVAR and 3D-LBP. Also, the width of the Parzen-Window kernel function, h, was fixed at 1.5 voxels in both ETDS algorithms.

- **Voxel Intensity Features** – The pyramid representation of the original volume was expanded up to level $l_{max} = 4$ and, for each layer, a binary brain mask was constructed and applied to exclude extracranial voxels which contain no information.

- **3D Local Variance** – Since these features have the same domain as VI (level $l = 0$), i.e., each feature is associated with one brain location, the same brain mask was used to retain only intracranial positions. As for LVAR parameters, R and P, features from three different configurations $(R, P) \in \{(2, 98), (4, 390), (6, 870)\}$ were pooled to form the feature vector.

- **Local Binary Patterns** – Three different combinations of parameters were considered: $(R, P) \in \{(2, 16), (4, 32), (6, 48)\}$ for 2D LBPs and $(R, P) \in \{(2, 24), (4, 24), (6, 24)\}$ for 3D LBPs. Moreover, the cubic mesh used to extract this type of feature was tuned by varying the cube dimension a. The following values were considered: $a \in \{9, 13, 17, 21, 25, 29, 33\}$. In addition, only histograms computed within cubes whose center lay inside the brain were used.

- **SVM Kernel** – The SVM was trained with both linear and RBF kernel functions. When the linear kernel was used, $C \in \left\{2^{-16}, 2^{-14}, 2^{-12}, 2^{-10}, 2^{-8}, 2^{-6}, 2^{-4}, 2^{-2}, 2^{0}, 2^{2}, 2^{4}\right\}$. For the RBF kernel, both C and γ were allowed to assume one of the following values: $\{2^{-18}, 2^{-14}, 2^{-10}, 2^{-6}, 2^{-2}, 2^{2}, 2^{6}, 2^{10}\}$.

- **Nested Cross-Validation** – The dataset was partitioned into 10 groups both in the inner and outer loop ($k = k' = 10$).

6.5 Classification Results

Before proceeding to the exposition of the performances attained by all implemented classifiers, it should be noted that no statistical hypothesis tests will be used in the current work for comparison purposes, and all comparisons will be based on the average and standard deviation of the performance measures. In fact, the McNemar's statistical test, which is the one used to compare results from two classifiers based on the same dataset and assuming a binomial distribution for the proportion of correctly classified subjects, was implemented but, due to the low number of training instances, this test did not found sufficient evidence to decide which methods are best, even when large differences in the accuracies were present.

The remainder of the current chapter will be focused on the assessment of all proposed systems for the CAD of AD. The procedures explored for selection purposes will be compared in section 6.5.1 and the different types of feature will be compared in section 6.5.2. All results will be listed in the summary Tables 6.6 and 6.7, presented in section 6.6.

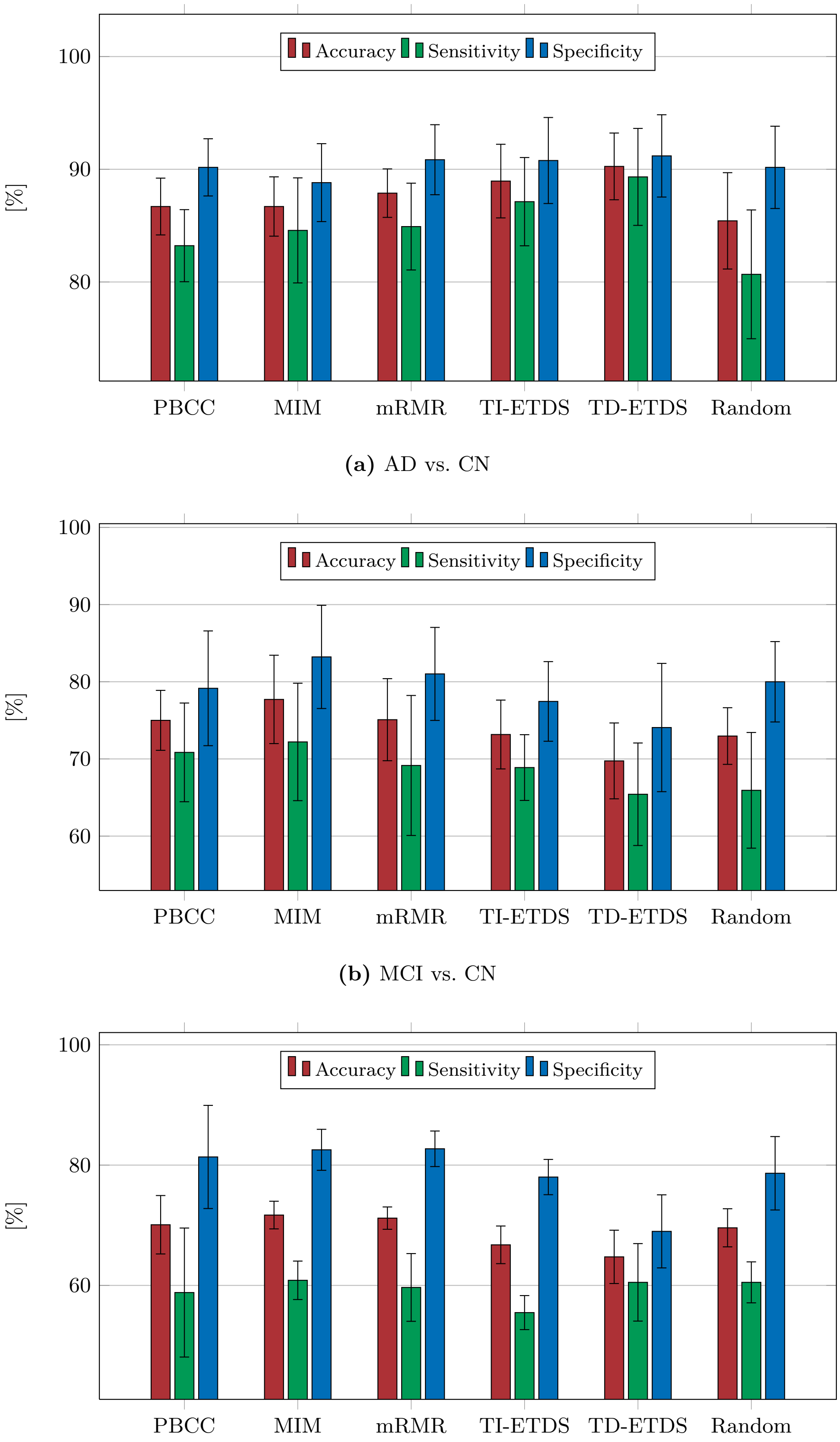
6.5.1 Feature Selection Algorithms

Chapter 4 presented five techniques to select, hopefully, the most discriminant features, and discussed their main advantages and disadvantages. Herein, their performances will be compared using the original VI features and the SVM algorithm with a linear kernel to learn all three classification tasks.

Figures 6.6(a), 6.6(b) and 6.6(c) compare the mean accuracies, sensitivities and specificities, presenting as well the corresponding error bars which represent the $\mu \pm 2\sigma$ interval. The statistical dispersion of these three quantities can have two independent sources: the random partitioning of the input data in the nested CV procedure and the random feature selection carried by both eye tracking algorithms.

Regarding the AD vs. CN task (Figure 6.6(a)), medically driven selection procedures seem to attain slightly better results than fully automated algorithms with the best marks (90.3% acc., 89.3% sens. and 91.2% spec.) being achieved by TD-ETDS. Nevertheless, all selection procedures achieved similar accuracies and one can not conclude with certainty if any of the studied algorithms is clearly superior or inferior to the others. When the intermediate state MCI is involved, the performances drop substantially, as already anticipated. MIM achieved the best accuracy both for MCI vs. CN (77.7%) (Figure 6.6(b)), and for AD vs. MCI (71.7%) (Figure 6.6(c)). On the opposite side, for the previous two classification tasks, both algorithms based on eye tracking data performed worse than all other procedures including random selection.

The power of each selection procedure can be better assessed for small number of features since, as this number increases, the effect of selection fades out, reason why even random selection behaved relatively well in all problems. However, the aforementioned results were computed using the nested CV

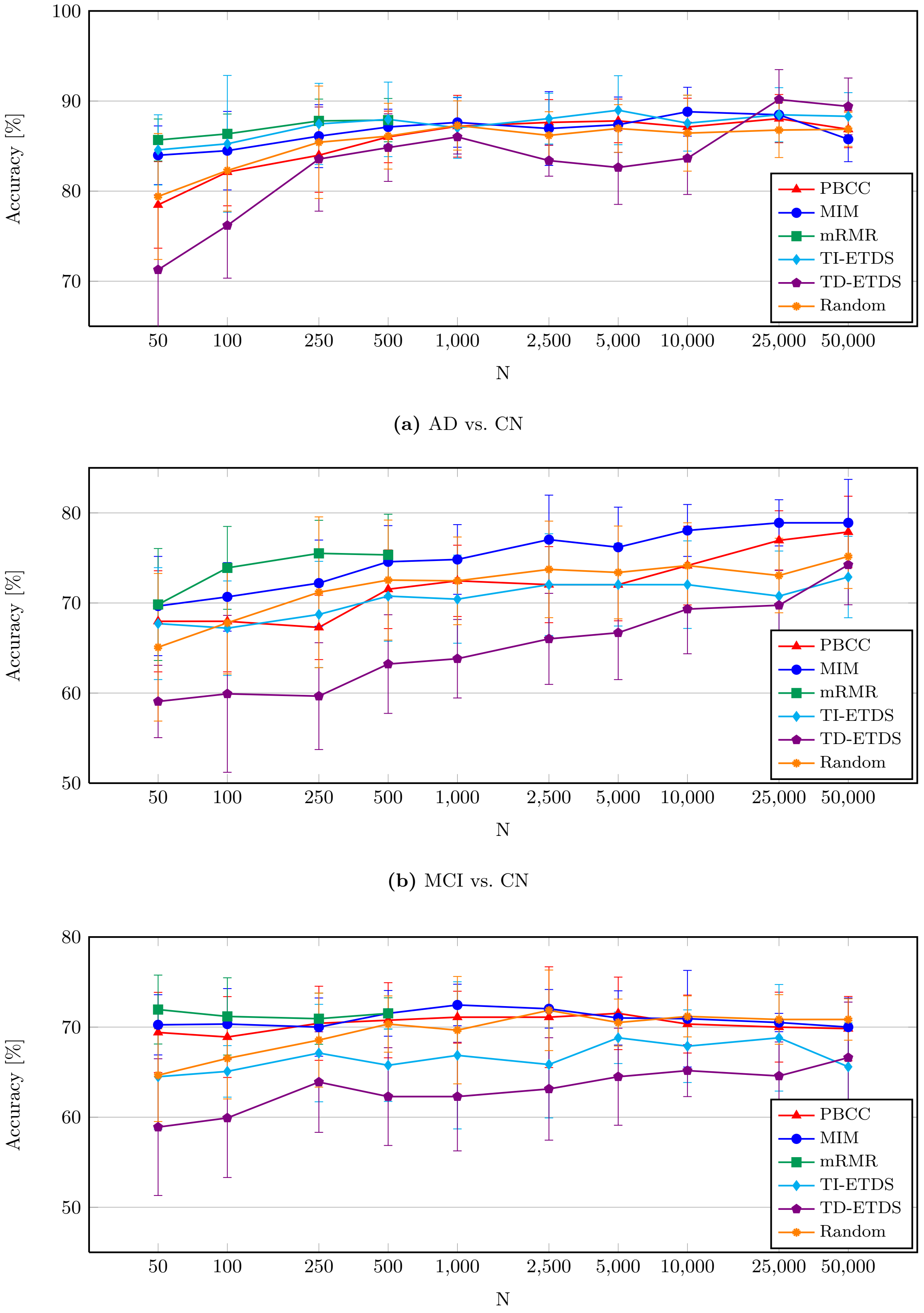


(a) AD vs. CN

(b) MCI vs. CN

(c) AD vs. MCI

Figure 6.6: Mean accuracies, sensitivities and specificities achieved by each feature selection technique using VI features for the three classification tasks. An error bar is illustrated for each result, representing the interval $\mu \pm 2\sigma$.



(a) AD vs. CN

(b) MCI vs. CN

(c) AD vs. MCI

Figure 6.7: Comparison of feature selection algorithms for varying values of N. VI features and the linear SVM kernel were used. Mean accuracies are presented as well as the two standard deviation interval.

procedure, meaning that each testing partition can be evaluated using a model built over a different set of features, and even with different values of N. In order to observe the influence of this parameter on the classification results, one nested CV for each value of N was performed. The results are presented in Figures 6.7(a), 6.7(b) and 6.7(c) for AD vs. CN, MCI vs. CN and AD vs. MCI, respectively. Note the nested CV is used to tune the SVM kernel parameter C.

The analysis of Figure 6.7 reveals important tendencies. First, for AD vs. CN, Figure 6.7(a) shows surprisingly that TD-ETDS, the method that achieved the best performance, is actually the only one performing significantly worse than random selection for feature spaces of dimension 10000 or lower. In fact, apart from PBCC and TD-ETDS, all other algorithms outperformed this dummy procedure, indicating that discriminative features are being selected first. This observation is further supported by the fact that most procedures do not seem to benefit much with the increasing number of selected features, except for PBCC, TD-ETDS and random selection. In the other two classifications tasks, MCI vs. CN and AD vs. MCI, Figures 6.7(b) and 6.7(c) (respectively) show that both medically driven procedures have difficulties in choosing the best VI features and therefore performed consistently worse than most selection algorithms for all values of N.

On a different note, mRMR outperformed all methods in all diagnostic problems for corresponding number of chosen voxels, but since it could only be evaluated up to 500 features, its marks were always exceeded in higher dimensional spaces. This is consistent with its theoretical advantage over the other studied methods. In fact, since mRMR avoids redundancy between selected features, this algorithm certainly joins a higher amount of information in a feature set of a given size.

Finally, it is important to notice that, apart from a few exceptions, the increase of N did not deteriorate the results and was even able to improve the performance of all classifiers in the task MCI vs. CN, proving that SVM was able to circumvent the *curse of dimensionality*. In order to see if classification performances continued to improve or remained stable even for more than 50000 features, the system was tested once using all 319441 available features, attaining 85,8%, 73,6% and 70.0% acc. for AD vs. CN, MCI vs. CN and AD vs. MCI, respectively. These marks are worse than the best results obtained in lower dimensional spaces, indicating that although the SVM algorithm alleviated the difficulties related with high dimensional spaces, the increase of N will eventually jeopardize the performance of the CAD system and, therefore, good selection procedures are essential to achieve the best results possible.

6.5.2 Feature Extraction Algorithms

The second goal of the present thesis was to assess alternative feature extraction techniques. In chapter 3, three types of features were introduced, VI, LVAR and LBP, and their classification results will now be presented. In order to maximize the performance of each feature type, three selection algorithms (PBCC, MIM and mRMR) were tested, as well as two types of kernel (Linear and RBF). In addition, for VI features, all layers of the scale-space were also tried. As stated before, since it was too time-demanding to evaluate all combinations of feature extraction, feature selection and kernel type, the best accuracy was searched step-by-step.

First of all, each level of the PET image's scale-space, from the original input image ($l = 0$) up to the one with the lowest resolution ($l = 4$), was tested using a linear kernel and MIM as the feature selection technique. The mean accuracies are displayed in Table 6.5, which shows not only the results obtained for different values of N, i.e., using the nested CV to optimize only the SVM parameter C, but also, in the last row of each table, the unbiased accuracy estimation reported by the nested CV procedure when used to tune both N and C.

Two important observations should be drawn from Table 6.5. On one hand, lowering the image resolution seems to enhance the performance in low dimensional feature spaces, but only up to a certain point after which the observed accuracies dropped abruptly. On the other hand, for the MCI vs. CN task, the overall result was actually improved by the usage of a lower resolution level ($l = 1$), achieving 79.4% accuracy, more 3.9% than in level $l = 0$. In fact, even for the other two classification tasks, the second layer ($l = 1$) always scored very close to the first one, achieving the same result for AD vs. CN and -0.6% for AD vs. MCI. Bearing this in mind, and also that the number of features in the second layer is 8 times smaller than in the first one, which represents a significant speed up of the learning stage of the CAD tool, level $l = 1$ was chosen to represent the VI features in the remainder of this work.

In the second stage, the best feature selection procedure was sought for each feature extraction alternative and in each classification task, by testing all three fully automated procedures: PBCC, MIM and mRMR. The same kernel was used in all experiments: the linear one. The best results are summarized in Figure 6.8.

MIM was most frequently the best algorithm (ranking first in 7 out of 12 problems), followed by PBCC which ranked first in 4 comparisons, and finally mRMR which got the best results in just one problem. In fact, PBCC and MIM achieved very similar results in most classifications, contrarily to mRMR which performed significantly poorer in several occasions. It should be stressed however that the possible number of features used in this algorithm had to be severely reduced in order to be able to produce results in an acceptable amount of time. More precisely, for the LVAR and 3D-LBP types of feature, N was only allowed to go as high as 100, while for the other types 500 was the maximum, a much smaller number when compared to the possible 50000 features selected by PBCC and MIM. In addition, although not presented here due to space limitations, each system was also evaluated with a fixed number of features in the nested CV procedure, in order to study the influence of N. Consistently with what was observed in Figure 6.7 for VI features, mRMR outperformed PBCC and MIM in several occasions, in low dimensional spaces.

Finally, the RBF kernel was tested for each type of feature using the settings chosen so far, i.e., the layer $l = 1$ of the scale-space when the VI features are involved and the best feature selection procedures found in the previous step. Figure 6.9 offers a clear comparison between the performance attained when using the linear and the RBF kernels. The usage of the RBF kernel did not improve significantly the performance of any type of feature, achieving similar or worse accuracies in all settings, despite the learning stage being much more time consuming due to the number of parameters to optimize.

Figure 6.9 also holds the best results achieved in the present study for each classification task. The

Table 6.5: Classification accuracy obtained for different layers of the scale-space using MI as selection criterion and a linear SVM kernel for varying number of selected features. The last row of each table also shows the accuracy achieved using the nested CV procedure to tune both N and C parameters. The best results obtained in each row are marked in boldface type.

AD vs. CN

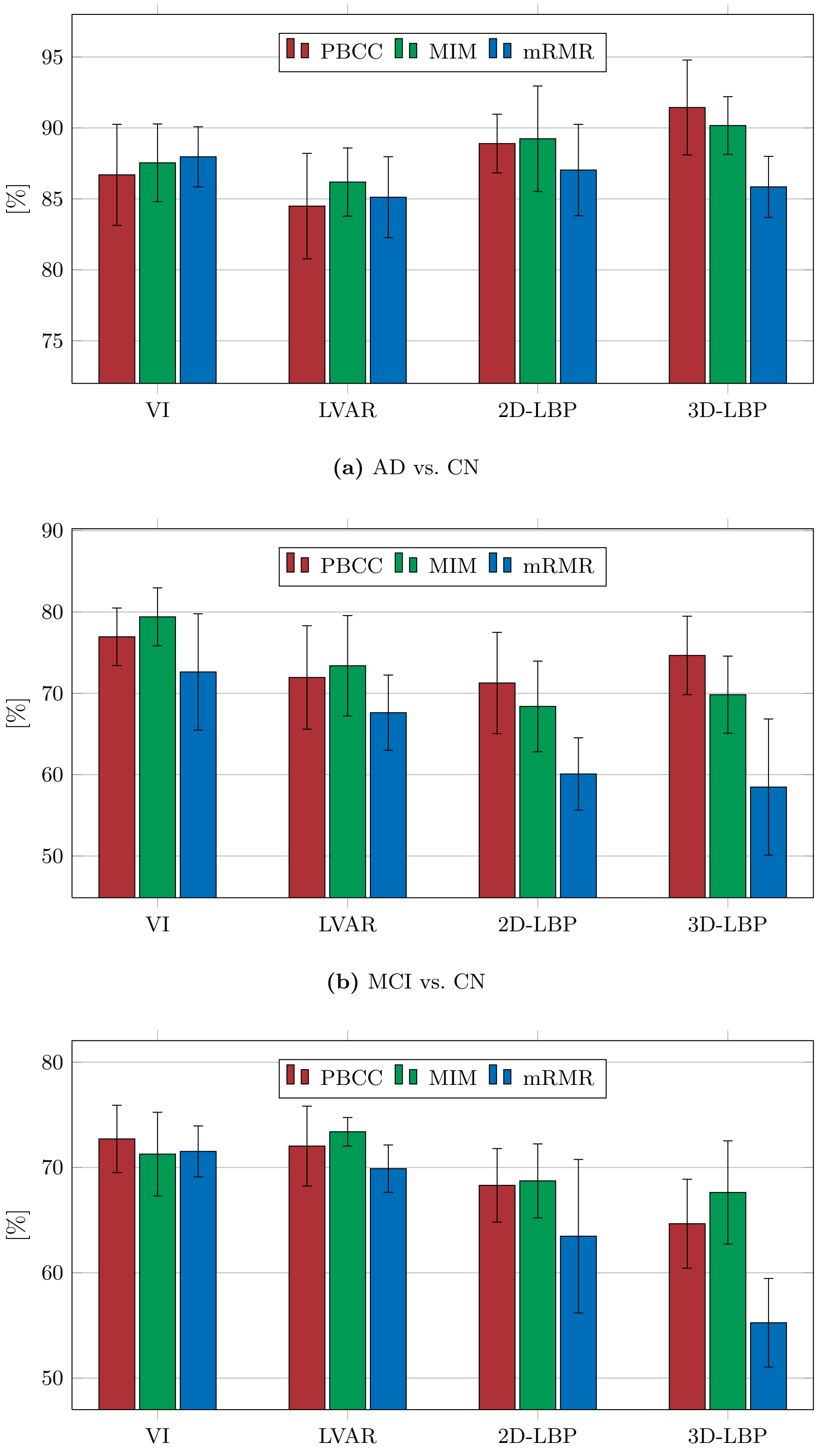
N	Level 0	Level 1	Level 2	Level 3	Level 4
50	**84.0**	82.9	82.8	81.5	71.9
100	84.5	85.1	**86.6**	84.4	72.8
250	86.1	**87.5**	85.8	84.0	77.7
500	87.1	**87.3**	86.1	83.6	
1000	87.6	**88.6**	86.5	83.6	
2500	86.9	**88.5**	86.2		
5000	**87.4**	86.9	86.3		
10000	**88.8**	86.4			
25000	**88.5**	87.8			
50000	**85.8**				
Accuracy	**87.5**	**87.5**	85.4	84.2	74.6

MCI vs. CN

N	Level 0	Level 1	Level 2	Level 3	Level 4
50	69.7	70.7	68.0	**71.7**	59.1
100	70.7	71.8	**73.0**	72.6	63.2
250	72.2	73.6	**75.4**	71.8	63.3
500	74.6	74.0	**75.3**	71.4	
1000	74.8	**78.2**	74.5	70.1	
2500	77.0	**80.3**	74.2		
5000	76.2	**80.7**	75.2		
10000	78.1	**79.7**			
25000	**78.9**	78.1			
50000	**78.9**				
Accuracy	75.5	**79.4**	74.7	71.3	62.1

AD vs. MCI

N	Level 0	Level 1	Level 2	Level 3	Level 4
50	70.3	70.8	**71.5**	68.4	66.3
100	70.3	**71.4**	70.9	67.7	65.1
250	70.0	71.3	**71.4**	66.5	66.2
500	71.5	**72.3**	70.3	66.7	
1000	**72.5**	71.8	70.6	66.1	
2500	**72.0**	70.7	70.5		
5000	**71.0**	70.0	67.2		
10000	**70.9**	70.3			
25000	**70.5**	70.3			
50000	**70.0**				
Accuracy	**71.9**	71.3	70.8	67.5	66.5

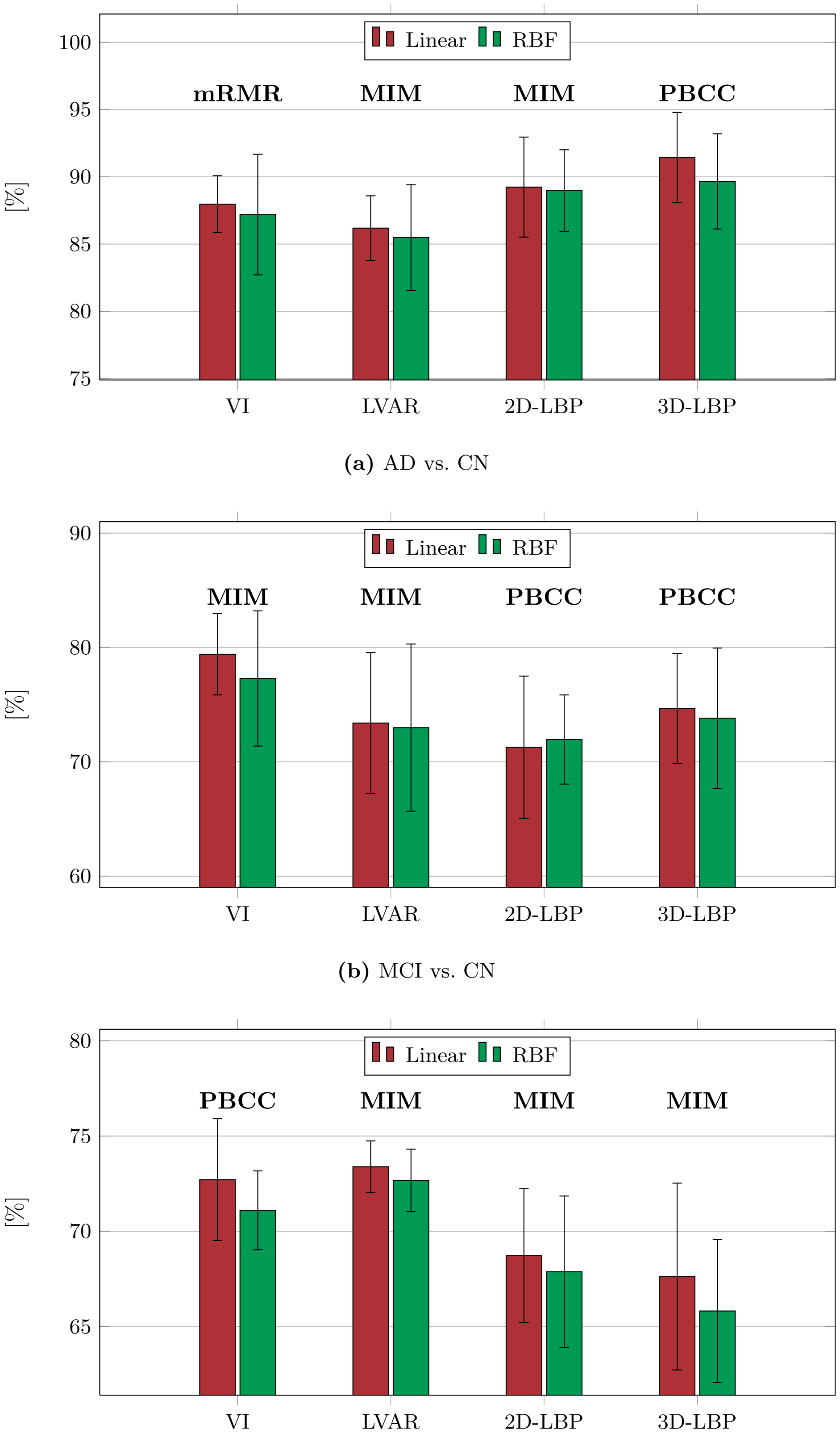


(a) AD vs. CN

(b) MCI vs. CN

(c) AD vs. MCI

Figure 6.8: Mean accuracies achieved by each selection algorithm for each type of feature. These results were obtained with a linear kernel and, once again, the two standard deviation interval is shown.



(a) AD vs. CN

(b) MCI vs. CN

(c) AD vs. MCI

Figure 6.9: Comparison between mean accuracies achieved by each type of feature, using the best selection algorithm (mentioned in the chart), and varying the kernel type. The error bars represent the two standard deviation interval.

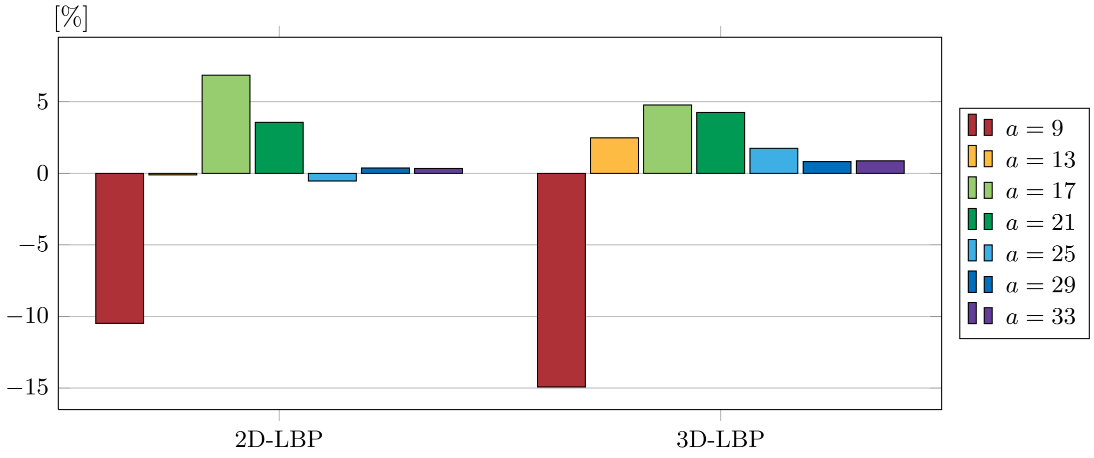
novel extension of LBP to three dimensions achieved the best results (91.4% acc., 90.5% sens. and 92.4% spec.) for the AD vs. CN task using PBCC to perform feature selection and an SVM with linear kernel for classification purposes. As regards MCI vs. CN, the best performance was attained by the level $l = 1$ of the scale-space of FDG-PET images together with MIM and linear kernel, yielding 79.4% acc., 75.9% sens. and 82.9% spec.. Finally, the combination of LVAR, MIM and a linear kernel outperformed all other settings in the classification of AD vs. MCI, reaching 73.4% acc., 62.9% sens. and 83.9% spec.. Additionally, the analysis of Tables 6.6 and 6.7 also reveals another interesting fact: specificity yielded consistently better results than sensitivity in all three classification tasks, meaning that a negative classification (i.e., CN in the tasks AD vs. CN and MCI vs. CN or MCI in the task AD vs. MCI) is more reliable than a positive one.

On a final subject, in order to better understand the influence of the parameters attached to the feature extraction procedures, namely, the cardinality of the neighbor set P and the radius R of the sphere where neighbors lie, used both for LVAR and LBPs, and the size a of the cube used only for LBPs, the percentage of occurrence of features originated by each combination of parameters was computed and compared before and after selection. In this comparison only instances from the AD and CN groups were considered and the selection technique that better separated those classes was used to select 1% of the total number of initially available features. The difference in the percentage of occurrence before and after the selection is shown in Figure 6.10.

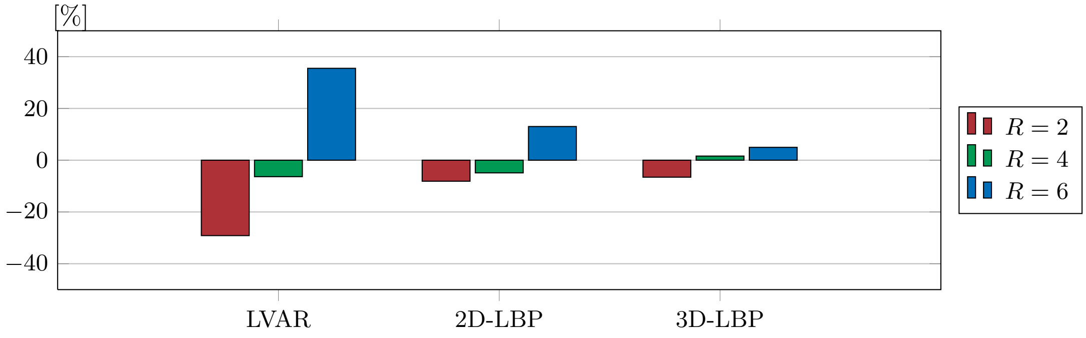
Its analysis leads to two conclusions. On one hand, it seems that the preferred LBP features are extracted from histograms computed for intermediate values of a. This is consistent with two observations previously made. First, smaller values of a means that LBP features are extracted from histograms using a smaller number of LBP instances, increasing the uncertainty attached to the estimation of the incidence rate of different LBP labels. As a consequence, if a given LBP's incidence rate contains useful information about the classification problem, this uncertainty can only jeopardize the feature's separation ability. Second, larger values of a will result in a coarser mesh and each histogram will merge information from larger areas of the brain, and eventually from regions characterized by more than one texture, therefore reducing their discriminative power. It should be noted however that the number of features selected with $a = 9$ or $a = 13$ is still very high because, initially, they were the most predominant ones. On the other hand, large values of R are preferably chosen, both for LVAR and LBPs. This tendency is most noticeable for LVAR where features created from a farther neighborhood showed a selection increase of almost 40%.

6.6 Summary

The results of all implemented classifiers are summarized in this section. Table 6.6 contains the mean accuracy, sensitivity and specificity obtained with VI features for the three classification tasks studied in this work. Table 6.7 presents the mean results using the other three types of feature: LVAR, 2D-LBP and 3D-LBP.



(a) Parameter a



(b) Parameter R

Figure 6.10: Difference in the percentage of occurrence of features produced by each combination of parameters between the following two situations: considering just 1% of the number of initially available features retained by the best selection algorithm in the AD vs. CN problem; considering all available features.

Table 6.6: Mean performances for each trained classifier – Part I. The best accuracy, specificity and sensitivity obtained in each classification task are presented in boldface type. The “Mean(Standard Deviation)” format is used.

Feature Type	Selection Algorithm	SVM Kernel	AD vs. CN			MCI vs. CN			AD vs. MCI		
			% ACC	% SENS	% SPEC	% ACC	% SENS	% SPEC	% ACC	% SENS	% SPEC
VI ($l = 0$)	PBCC	Linear	86.7(1.3)	83.2(1.6)	90.2(1.3)	75.0(1.9)	70.8(3.2)	79.2(3.7)	70.1(2.4)	58.8(5.4)	81.4(4.3)
	MIM	Linear	87.5(1.8)	85.9(2.6)	89.2(1.4)	75.5(1.8)	69.8(3.1)	81.2(1.9)	71.9(1.6)	60.8(1.6)	82.9(2.9)
	mRMR	Linear	87.9(1.1)	84.9(1.9)	90.8(1.6)	75.1(2.7)	69.2(4.5)	81.0(3.0)	71.2(0.9)	59.7(2.8)	82.7(1.5)
	TI-ETDS	Linear	89.0(1.6)	87.8(2.0)	90.2(1.9)	72.9(2.3)	69.4(2.1)	76.4(2.6)	68.8(2.6)	58.7(1.4)	78.9(2.8)
	TD-ETDS	Linear	90.3(1.5)	89.3(2.2)	91.2(1.8)	69.7(2.5)	65.4(3.3)	74.1(4.2)	64.7(2.2)	60.5(3.2)	69.0(3.0)
	Random	Linear	85.4(2.1)	80.7(2.9)	90.2(1.8)	73.0(1.8)	65.9(3.7)	80.0(2.6)	69.6(1.6)	60.5(1.7)	78.6(3.1)
VI ($l = 1$)	PBCC	Linear	86.7(1.8)	83.2(3.3)	90.2(1.5)	76.9(1.8)	72.4(2.9)	81.5(1.9)	72.7(1.6)	63.7(3.1)	81.7(1.8)
		RBF	—	—	—	—	—	—	71.1(1.0)	63.2(2.5)	79.0(1.6)
	MIM	Linear	87.5(1.4)	85.4(2.4)	89.7(1.2)	**79.4(1.8)**	**75.9(3.0)**	**82.9(1.9)**	71.3(2.0)	62.4(3.3)	80.2(2.4)
		RBF	87.2(2.2)	85.1(3.3)	89.3(1.8)	77.3(3.0)	73.2(3.9)	81.4(3.2)	–	–	–
	mRMR	Linear	88.0(1.1)	84.9(1.8)	91.0(1.7)	72.6(3.6)	66.9(3.1)	78.3(5.1)	71.5(1.2)	60.7(3.7)	82.4(2.6)
VI ($l = 2$)	MIM	Linear	85.4(2.1)	82.7(2.9)	88.1(2.1)	74.7(3.4)	70.2(5.0)	79.3(3.6)	70.8(2.0)	61.7(2.0)	79.8(2.5)
VI ($l = 3$)	MIM	Linear	84.2(2.4)	81.0(2.4)	87.3(3.1)	71.3(2.0)	69.2(4.1)	73.4(3.4)	67.5(1.9)	59.0(3.1)	75.9(3.4)
VI ($l = 4$)	MIM	Linear	74.6(2.0)	71.5(3.2)	77.6(1.8)	62.1(2.1)	61.4(3.0)	62.9(3.3)	66.5(1.7)	59.8(2.4)	73.2(2.7)

Table 6.7: Average performances for each trained classifier – Part II. The best accuracy, specificity and sensitivity obtained in each classification task are presented in boldface type. The “Mean(Standard Deviation)” format is used.

Feature Type	Selection Algorithm	SVM Kernel	AD vs. CN			MCI vs. CN			AD vs. MCI		
			% ACC	% SENS	% SPEC	% ACC	% SENS	% SPEC	% ACC	% SENS	% SPEC
LVAR	PBCC	Linear	84.5(1.9)	79.0(2.6)	90.0(2.7)	71.9(3.2)	69.5(4.4)	74.4(4.6)	72.0(1.9)	62.4(2.4)	81.7(2.6)
	MIM	Linear	86.2(1.2)	81.2(1.9)	91.2(2.1)	73.4(3.1)	72.4(1.9)	74.4(5.3)	**73.4(0.7)**	62.9(1.2)	83.9(1.1)
		RBF	85.5(2.0)	81.1(3.5)	89.8(3.4)	73.0(3.7)	68.9(3.7)	77.1(5.4)	72.7(0.8)	59.3(1.7)	**86.0(1.8)**
	mRMR	Linear	85.1(1.4)	82.2(2.9)	88.0(1.6)	67.6(2.3)	62.3(3.8)	72.9(3.3)	69.9(1.0)	63.4(1.7)	76.4(1.4)
2D-LBP	PBCC	Linear	88.9(1.0)	83.7(3.1)	94.1(2.5)	71.3(3.1)	68.8(4.2)	73.7(5.4)	68.3(1.7)	58.5(2.9)	78.1(3.2)
		RBF	—	—	—	71.9(1.9)	69.7(3.4)	74.2(3.0)	—	—	—
	MIM	Linear	89.2(1.9)	83.4(2.8)	**95.1(2.7)**	68.4(2.8)	66.8(2.6)	70.0(4.1)	68.7(1.8)	59.5(3.2)	78.0(4.3)
		RBF	89.0(1.5)	84.2(2.3)	93.7(2.3)	—	—	—	67.9(2.0)	61.9(2.5)	73.9(3.3)
	mRMR	Linear	87.0(1.6)	79.8(3.2)	94.2(1.7)	60.1(2.2)	55.4(4.7)	64.7(3.3)	63.5(3.6)	47.8(3.5)	79.2(4.8)
3D-LBP	PBCC	Linear	**91.4(1.7)**	**90.5(3.4)**	92.4(2.3)	74.7(2.4)	69.0(3.9)	80.3(2.2)	64.7(2.1)	57.5(2.1)	71.9(3.6)
		RBF	89.7(1.8)	88.0(3.3)	91.4(1.9)	73.8(3.1)	71.5(4.8)	76.1(4.3)	—	—	—
	MIM	Linear	90.2(1.0)	86.4(1.5)	93.9(1.7)	69.8(2.4)	63.2(3.1)	76.4(4.1)	67.6(2.5)	**63.9(4.2)**	71.4(4.1)
		RBF	—	—	—	—	—	—	65.8(1.9)	61.4(3.9)	70.2(3.6)
	mRMR	Linear	85.8(1.1)	79.7(2.7)	92.0(1.9)	58.5(4.2)	49.0(5.9)	68.0(4.4)	55.3(2.1)	47.6(2.8)	62.9(4.4)

7

Conclusions and Future Work

The current thesis studied several approaches for the automatic classification of Alzheimer's disease based on FDG-PET images. Nowadays, the interpretation of brain images, such as the one just mentioned, MRI or SPECT, still depends completely on the physician expertise and, as a consequence, there has been a growing interest in developing methods to reliably distinguish people suffering from AD and related disorders from normal controls. In addition, the existence of increasingly discriminant biomarkers has enabled the diagnosis of AD in its early stages. Therefore, in this work, besides the AD vs. CN classification task, the MCI state, a medical condition which often precedes the onset of AD, was also considered and two more classifications tasks were studied, MCI vs. CN and AD vs. MCI.

The large number of voxels in brain images together with the comparatively small number of instances available for learning purposes poses a problem commonly known as the *curse of dimensionality*, and which often jeopardizes the performance of any pattern recognition system. To tackle this problem, several feature selection algorithms were studied in order to reduce significantly the dimensionality of the feature space, while trying to retain as much information as possible.

Two medically driven selection algorithms were considered: TI-ETDS and TD-ETDS. The latter (TD-ETDS) is, in fact, a novel extension to the former (TI-ETDS), capable of mimicking an expert physician not only in the choice of the most important voxels but also in the comparison of different regions of the brain. The innovative approach TD-ETDS achieved the best results in the classification of AD vs. CN with an accuracy of 90.3%, but when the MCI state was involved, the reported results were much worse, achieving worse performances even when compared to random selection, which makes no effort to find the best features. Therefore, one can conclude that in both MCI vs. CN and AD vs. MCI tasks, TD-ETDS seems to overlook the most important regions of the FDG-PET image. The same behavior was found for the TI-ETDS procedure. The lower performance of both ETDS techniques may be related to the fact that eye tracking data was recorded while the physician was performing multi-class classification (CN vs. MCI vs. AD), while here we are focused on dichotomous classification problems. Nevertheless, in the classification of AD vs. CN, both algorithms performed better than most expert physicians [83] and than most classifiers found in the literature (see Tables 2.1 and 2.2) and can still be improved. For instance, the inclusion of eye tracking data from more than one physician would certainly help to build a more reliable model, as well as the use of data acquired from a physician faced with dichotomous problems.

In addition, three fully automated feature selection procedures were also tested. Beyond the ranking algorithms, PBCC and MIM, commonly used in this research field, mRMR was also implemented which, despite being a recognized selection procedure, had never been used before for the CAD of AD, perhaps due to its higher computational requirements. However, the study of a selection algorithm that takes into account the redundancy between selected features was considered to be very relevant in this specific problem, due to the high correlation nature of neighboring voxels in the FDG-PET image. Regarding classification results, MIM reported the best performance for VI features (the type of features used to compare all selection algorithms) in both MCI vs. CN and AD vs. MCI with accuracies of 75.5% and 71.9%, respectively. Also, even when other types of features were being used, MIM was often the best method, confirming in practice its theoretical advantage over PBCC. The use of mRMR led to inferior performances in all settings but one. However, that was probably only motivated by the low number of features that this algorithm can select in an acceptable amount of time, since it was shown that mRMR was consistently better in low dimensional spaces, indicating that better performances might be achieved if it was possible to consider connections between features at lower computational costs. One possibility is to abandon the paradigm of feature selection and consider algorithms that project the data onto low dimensional spaces such as LDA or PCA. The computational burn of these methods is smaller since they look for an optimal linear combination of the input features, i.e., the optimal coefficients in a continuous problem, while feature selection algorithms can be seen as solvers of a discrete optimization problem, where the low dimensional space is obtained through the linear combination of features, with the coefficients restricted to the values 0 and 1.

The second objective of the current study was to evaluate the use of features of nature different than the raw voxel intensities. In this regard, two feature extraction algorithms were tested: LVAR which measures for each brain's position the contrast of a small neighborhood, and the widely known texture descriptor LBP. In addition, FDG-PET images were also expanded in their scale-space representation in order to study the effect of the image's resolution. The performances achieved in all classification tasks were improved by some of these transformations, specifically, AD vs. CN was best classified by 3D-LBPs with 91.4% accuracy, MCI vs. CN by the second level of the scale-space with 79.4% accuracy and AD vs. MCI by LVAR with 73.4% accuracy. This study also showed that the loss incurred by reducing the resolution of the original input images was negligible for a subsampling factor of 8 and it even enhanced the performance for the MCI vs. CN classification task, which is highly significant considering the substantial decrease in the starting number of features and the corresponding computational gain. As regards the LBP type of feature, a novel approach to the extension of the original extraction algorithm to three-dimensional data was proposed. This extension differentiates itself from others found in the literature by not introducing any approximation to the original concepts. In addition, three-dimensional LBP achieved good overall performances, improving the results of its two-dimensional counterpart in the AD vs. CN and MCI vs. CN tasks.

In the context of feature extraction several changes might also improve the overall performance. The recent discovery of the new substance, Pittsburgh compound B (PiB), will certainly change the course of this research field in the near future, since when used as tracer in PET technology, it becomes possible to detect the presence of beta-amyloid plaques in the brain tissue, one of the hallmarks of Alzheimer's disease. In addition, the number of PiB-PET scans available, for instance, at the ADNI database is already large enough to apply the methods herein described. Moreover, the integration of multiple biomarkers represents one more possible direction to the current study.

Finally, the present work benefited from the robustness of the SVM algorithm to almost empty spaces, which alleviated the *curse of dimensionality.* In fact, SVM was vital to achieve very good performances using feature vectors of dimensionality as high as 50000 features and only 118 training instances. In what concerns the learning stage of the CAD system, a natural follow-up work is to merge the three dichotomous problems and perform multi-class classification or even to introduce scans from patients suffering from other types of dementia in order to come closer to a real life environment.

Bibliography

[1] R. Brookmeyer, S. Gray, and C. Kawas, "Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset," *American Journal of Public Health*, vol. 88, no. 9, pp. 1337–1342, 1998.

[2] A. Association, "2012 Alzheimer's disease facts and figures," *Alzheimer's and Dementia: The Journal of the Alzheimer's Association*, vol. 8, no. 2, pp. 131–168, 2012.

[3] C. P. Ferri, R. Sousa, E. Albanense, W. s. Ribeiro, and M. Honyashiki, "World Alzheimer Report 2009," 2009.

[4] A. Wimo and M. Prince, "World Alzheimer Report 2010: The global economic impact of dementia," September 2010.

[5] A. Europe, "2006: Dementia carers' survey," 2006.

[6] R. Luengo-Fernandez, J. Leal, and A. Gray, "Dementia 2010: The economic burden of dementia and associated research funding in the United Kingdom," 2010.

[7] A. Drzezga, "Concept of functional imaging of memory decline in Alzheimer's disease," *Methods*, vol. 44, no. 4, pp. 304–314, 2008.

[8] D. Silverman, *PET in the Evaluation of Alzheimer's Disease and Related Disorders*. Springer, 2009.

[9] D. H. S. Silverman, "Brain 18F-FDG PET in the diagnosis of neurodegenerative dementias: comparison with perfusion SPECT and with clinical evaluations lacking nuclear imaging," *Journal of Nuclear Medicine*, vol. 45, no. 4, pp. 594–607, 2004.

[10] K. Herholz, S. Carter, and M. Jones, "Positron emission tomography imaging in dementia," *The British Journal of Radiology*, vol. 80, pp. 160–167, 2007.

[11] Laboratory of Neuro Imaging, UCLA, "About ADNI," http://adni.loni.ucla.edu/about/, 2012, [Online: accessed December 1, 2012].

[12] J. Stoeckel, G. Malandain, O. Migneco, P. M. Koulibaly, P. Robert, N. Ayache, and J. Darcourt, "Classification of SPECT images of normal subjects versus images of Alzheimer's disease patients," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI'01), Proceedings of the 4th International Conference on.* Springer-Verlag, pp. 666–674, 2001.

[13] J. Stoeckel and G. Fung, "SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information," *Knowledge and Information Systems*, vol. 11, no. 2, pp. 243–258, 2007.

[14] Y. Xia, L. Wen, S. Eberl, M. Fulham, and D. Feng, "Genetic algorithm-based PCA eigenvector selection and weighting for automated identification of dementia using FDG-PET imaging," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pp. 4812–4815, 2008.

[15] M. López, J. Ramírez, J. Górriz, D. Salas-Gonzalez, I. Álvarez, F. Segovia, and R. Chaves, "Multivariate approaches for Alzheimer's disease diagnosis using Bayesian classifiers," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, pp. 3190–3193, 2009.

[16] I. Illán, J. Górriz, J. Ramírez, D. Salas-Gonzalez, M. López, F. Segovia, R. Chaves, M. Gómez-Rio, and C. Puntonet, "18F-FDG PET imaging analysis for computer aided Alzheimer's diagnosis," *Information Sciences*, vol. 181, no. 4, pp. 903–916, 2011.

[17] M. Silveira and J. Marques, "Boosting Alzheimer disease diagnosis using PET images," in *Pattern Recognition (ICPR'10), Proceedings of the 2010 20th International Conference on.* IEEE Computer Society, pp. 2556–2559, 2010.

[18] E. Bicacro, M. Silveira, J. S. Marques, and D. C. Costa, "3D brain image-based diagnosis of Alzheimer's disease: Bringing medical vision into feature selection," in *Biomedical Imaging (ISBI'12), 9th IEEE International Symposium on*, pp. 134–137, 2012.

[19] K. R. Gray, R. Wolz, S. Keihaninejad, R. A. Heckemann, P. Aljabar, A. Hammers, and D. Rueckert, "Regional analysis of FDG-PET for use in the classification of Alzheimer's disease," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pp. 1082–1085, 2011.

[20] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.

[21] S. Duchesne, A. Caroli, C. Geroldi, C. Barillot, G. B. Frisoni, and D. L. Collins, "MRI-based automated computer classification of probable AD versus normal controls," *Medical Imaging, IEEE Transactions on*, vol. 27, no. 4, pp. 509–520, 2008.

[22] 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'12), 9th IEEE International Symposium on*, pp. 606– 609, 2012.

[23] K. Oppedal, K. Engan, D. Aarsland, M. K. Beyer, O.-B. Tysnes, and T. Eftestøl, "Using local binary pattern to classify dementia in MRI," in *Biomedical Imaging (ISBI'12), 9th IEEE International Symposium on.* IEEE, pp. 594–597, 2012.

[24] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, pp. 1226–1238, 2005.

[25] J. R. Petrella, R. E. Coleman, and P. M. Doraiswamy, "Neuroimaging and early diagnosis of Alzheimer disease: a look to the future," *Radiology*, vol. 226, no. 2, pp. 315–336, 2003.

[26] P. Vemuri, J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack, "Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies." *NeuroImage*, vol. 39, no. 3, pp. 1186–97, 2008.

[27] C. Akgül and A. Ekin, "A probabilistic information fusion approach to MR-based automated diagnosis of dementia," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 265–268, 2010.

[28] J. B. Fiot, J. Fripp, and L. D. Cohen, "Combining imaging and clinical data in manifold learning: Distance-based and graph-based extensions of Laplacian Eigenmaps," in *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pp. 570–573, 2012.

[29] J. M. Górriz, J. Ramirez, A. Lassl, D. Salas-Gonzalez, E. W. Lang, C. G. Puntonet, I. Alvarez, M. López, and M. Gómez-Rio, "Automatic computer aided diagnosis tool using component-based SVM," in *Nuclear Science Symposium Conference Record (NSS'08), IEEE*, pp. 4392–4395, 2008.

[30] R. Chaves, J. Ramírez, J. M. Górriz, M. López, I. Álvarez, D. Salas-Gonzalez, F. Segovia, and P. Padilla, "SPECT image classification based on NMSE feature correlation weighting and SVM," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, pp. 2715–2719, 2009.

[31] E. E. Tripoliti, D. I. Fotiadis, and M. Argyropoulou, "A supervised method to assist the diagnosis of Alzheimer's disease based on functional Magnetic Resonance Imaging," in *Engineering in Medicine and Biology Society (EMBS'07), 29th Annual International Conference of the IEEE*, pp. 3426–3429, 2007.

[32] E. Gerardin, G. Chetelat, M. Chupin, R. Cuingnet, B. Desgranges, H. Kim, M. Niethammer, B. Dubois, S. Lehericy, L. Garnero, F. Eustache, and O. Colliot, "Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging," *NeuroImage*, vol. 47, no. 4, pp. 1476–1486, 2009.

[33] E. Bicacro, M. Silveira, and J. S. Marques, "Alternative feature extraction methods in 3D brain image-based diagnosis of Alzheimer's disease," in *Image Processing (ICIP'12), 2012 IEEE International Conference on*, pp. 134–137, 2012.

[34] P. Padilla, M. López, J. M. Górriz, J. Ramirez, D. Salas-Gonzalez, and I. Álvarez, "NMF-SVM based CAD tool applied to functional brain images for the diagnosis of Alzheimer's disease," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 2, pp. 207–216, 2012.

[35] E. Bicacro, “Alzheimer’s disease diagnosis using 3D brain images: Exploring new grounds on features extraction and selection,” Master’s thesis, Instituto Superior Técnico, 2011.

[36] P. J. Burt, “Fast filter transform for image processing,” *Computer Graphics and Image Processing*, vol. 16, no. 1, pp. 20–51, 1981.

[37] ——, “Fast algorithms for estimating local image properties,” *Computer Vision, Graphics and Image Processing*, vol. 21, no. 3, pp. 368–382, 1983.

[38] P. J. Burt and E. H. Adelson, “The laplacian pyramid as a compact image code,” *Communications, IEEE Transactions on*, vol. 31, no. 4, pp. 532–540, 1983.

[39] E. Peli, “Contrast in complex images,” *Journal of the Optical Society of America A*, vol. 7, no. 10, pp. 2032–2040, 1990.

[40] S. Hill, *Tri-linear interpolation.* Academic Press, ch. 10.1, pp. 521–525, 1994.

[41] N. J. A. Sloan, R. H. Hardin, and W. D. Smith, “Table of spherical codes,” http://neilsloane.com/packings/, 2000, [Online: accessed December 1, 2012].

[42] R. H. Hardin, N. J. A. Sloan, and W. D. Smith, “Tables of spherical codes with icosahedral symmetry,” http://neilsloane.com/icosahedral.codes/, 2012, [Online: accessed December 1, 2012].

[43] T. Randen and J. H. Husøy, “Filtering for texture classification: A comparative study,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 4, pp. 291–310, 1999.

[44] L. G. Shapiro and G. C. Stockman, *Computer Vision.* Prentice Hall, ch. Texture, pp. 235–248, 2001.

[45] T. Mäenpää and M. Pietikäinen, *Texture analysis with local binary patterns.* World Scientific Publishing Company, ch. 11, pp. 197–216, 2005.

[46] H. Yang and Y. Wang, “A LBP-based face recognition method with hamming distance constraint,” in *Image and Graphics (ICIG 2007), Fourth International Conference on*, pp. 645–649, 2007.

[47] T. Ojala, M. Pietikäinen, and T. Mäenpää, “A comparative study of texture measures with classification based on featured distributions,” *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[48] ——, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[49] R. Jain, R. Kasturi, and B. Schunck, *Machine Vision.* McGraw-Hill, ch. Bilinear interpolation, pp. 382–383, 1995.

[50] G. Zhao and M. Pietikäinen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 915–928, 2007.

[51] J. Fehr, “Rotational invariant uniform local binary patterns for full 3D volume texture analysis,” in *Finnish Signal Processing Symposium (FINSIG’07), Proc.*, 2007.

[52] J. Fehr and H. Burkhardt, “3D rotation invariant local binary patterns,” in *Pattern Recognition (ICPR’08), 19th International Conference on*, pp. 1–4, 2008.

[53] D. Avis and B. Kaluzny, “Solving inequalities and proving Farka’s theorem made easy,” *American Mathematical Monthly*, vol. 111, pp. 152–157, 2004.

[54] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, “Rotation invariant spherical harmonic representation of 3D shape descriptors,” in *Geometry Processing (SGP ’03), Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on.* Eurographics Association, pp. 156–164, 2003.

[55] S. J. Raudys and A. K. Jain, “Small sample size effects in statistical pattern recognition: Recommendations for practitioners,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 252–264, 1991.

[56] S. J. Raudys and V. Pikelis, “On dimensionality, sample size, and classification error of nonparametric linear classification algorithms,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, pp. 667–671, 1997.

[57] A. K. Jain and B. Chandrasekaran, “Dimensionality and sample size considerations in pattern recognition practice,” in *Classification Pattern Recognition and Reduction of Dimensionality*, ser. Handbook of Statistics, P. Krishnaiah and L. Kanal, Eds. Elsevier, vol. 2, pp. 835–855, 1982.

[58] R. P. W. Duin, “Classifiers in almost empty spaces,” *Pattern Recognition, International Conference on*, vol. 2, pp. 1–7, 2000.

[59] I. Guyon and A. Elisseef, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[60] R. Jensen and Q. Shen, *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches.* Wiley-IEEE Press, 2008.

[61] J. L. Rodgers and W. Nicewander, “Thirteen ways to look at the correlation coefficient,” *The American Statistician*, vol. 42, pp. 59–66, 1988.

[62] C. E. Shannon, “A mathematical theory of communication,” Bell System technical journal, Tech. Rep. 27, 1948.

[63] R. Fano, *Transmission of Information: A Statistical Theory of Communications.* MIT Press, 1961.

[64] G. Brown, “A new perspective for information theoretic feature selection,” *Journal of Machine Learning Research*, vol. 5, pp. 49–56, 2009.

[65] G. Brown, A. Pocock, M. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[66] J. M. H.H. Yang, "Data visualization and feature selection: new algorithms for non-gaussian data," in *Advances in Neural Information Processing Systems*, vol. 12. MIT Press, pp. 687–693, 1999.

[67] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, 2004.

[68] A. T. Duchowski, *Eye Tracking Methodology: Theory and Practice*, ser. Methods in molecular biology. Springer, 2007.

[69] T. T. AB, "Toby in brief," http://www.tobii.com/en/group/about-tobii/tobii-in-brief/, 2010, [Online: accessed December 1, 2012].

[70] *Tobii T/X Series Eye Trackers*, 2nd ed., Toby Technology AB, 2010.

[71] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[72] V. Vapnik and A. Lerner, "Pattern recognition using Generalized Portrait method," *Automation and Remote Control*, vol. 24, 1963.

[73] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Computational Learning Theory (COLT '92), Proceedings of the 5th annual ACM workshop on.* ACM Press, pp. 144–152, 1992.

[74] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, vol. 20, pp. 273–297, 1995.

[75] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony, "Structural risk minimization over data-dependent hierarchies," *Information Theory, IEEE Transactions on*, vol. 44, no. 5, pp. 1926–1940, 1998.

[76] P. L. Bartlett, "The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network," *Information Theory, IEEE Transactions on*, vol. 44, no. 2, pp. 525–536, 2006.

[77] J. Shawe-Taylor and N. Cristianini, "Margin distribution and soft margin," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. MIT Press, pp. 349–358, 2000.

[78] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *Intelligent Systems and Technology, ACM Transactions on*, vol. 2, no. 3, pp. 1–27, 2011.

[79] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” *BMC Bioinformatics*, vol. 7, no. 1, 2006.

[80] Laboratory of Neuro Imaging, UCLA, “About the study,” http://adni.loni.ucla.edu/about/about-the-study/, 2012, [Online: accessed December 1, 2012].

[81] M. W. Weiner, P. S. Aisen, C. R. Jack, W. J. Jagust, J. Q. Trojanowski, L. Shaw, A. J. Saykin, J. C. Morris, N. Cairns, L. A. Beckett, A. Toga, R. Green, S. Walter, H. Soares, P. Snyder, E. Siemers, W. Potter, P. E. Cole, and M. Schmidt, “The Alzheimer’s disease neuroimaging initiative: progress report and future plans,” *Alzheimer’s and Dementia*, vol. 6, no. 3, pp. 202–211.e7, 2010.

[82] Laboratory of Neuro Imaging, UCLA, “PET protocols,” http://adni.loni.ucla.edu/research/protocols/pet-protocols/, 2012, [Online: accessed December 1, 2012].

[83] S. Klöppel, C. M. Stonnington, J. Barnes, F. Chen, C. Chu, C. D. Good, I. Mader, L. A. Mitchell, A. C. Patel, C. C. Roberts, N. C. Fox, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak, “Accuracy of dementia diagnosis – a direct comparison between radiologists and a computerized method,” *Brain*, vol. 131, no. 11, pp. 2969–2974, 2008.

[84] C. Müller, *Spherical harmonics: Lecture notes in mathematics.* Springer, vol. 17, 1966.

A

Spherical Harmonics

According to the theory of spherical harmonics, any square-integrable spherical function $f(\theta, \varphi)$ can be expanded as a linear combination of its harmonics [84]:

$$f(\theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_{l,m} Y_l^m(\theta, \varphi), \tag{A.1}$$

where $a_{l,m}$ is the complex coefficient associated with the spherical harmonic Y_l^m of degree l and order m which can be calculated as follows:

$$Y_l^m(\theta, \varphi) = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}} \cdot P_l^m(cos\theta) e^{im\varphi}, \tag{A.2}$$

where $P_l^m(\cdot)$ is the associated Legendre polynomial. In practice, the expansion (A.1) of a given function in its spherical harmonics needs to be truncated at a given maximum degree l_M. However, it is known that as l_M tends to infinity, the reconstruction tends to the true function [84]:

$$\lim_{l_M \to \infty} \sum_{l=0}^{l_M} \sum_{m=-l}^{l} a_{l,m} Y_l^m(\theta, \varphi) = f(\theta, \varphi), \tag{A.3}$$

or in other words, the error of the reconstruction tends to zero. In addition, spherical harmonics, whose real parts are illustrated in Figure A.1, form an orthonormal basis for square-integrable spherical functions [84].

Given a specific function $f(\theta, \varphi)$, the expansion coefficients $a_{l,m}$ can be obtained through its projection (inner product) into the respective base functions Y_l^m:

$$a_{l,m} = \langle f(\theta, \varphi), Y_l^m(\theta, \varphi) \rangle = \int_0^{2\pi} \int_0^{\pi} f(\theta, \varphi) Y_l^{m*}(\theta, \varphi) d\Omega, \tag{A.4}$$

where the superscript $*$ stands for the complex conjugate and Ω the solid angle ($d\Omega = \sin\theta d\theta d\varphi$).

The key property to the current work is that the subspace V_l formed by the span of the functions Y_l^m with l fixed to a given degree,

$$V_l = Span\left(Y_l^{-l}, Y_l^{-l+1}, \ldots, Y_l^{l-1}, Y_l^l\right), \tag{A.5}$$

is a representation for the rotation group, i.e. if a particular function f belongs to the subspace V_l, then any rotation $R(f)$ will stay in that subspace. In other words, let π_l be the projection onto the subspace V_l, then π_l commutes with rotations [54], i.e.:

$$\pi_l\left(R\left(f\right)\right) = R\left(\pi_l\left(f\right)\right), \tag{A.6}$$

where

$$\pi_l\left(f\left(\theta, \varphi\right)\right) = \sum_{m=-l}^{l} a_{l,m} Y_l^m(\theta, \varphi). \tag{A.7}$$

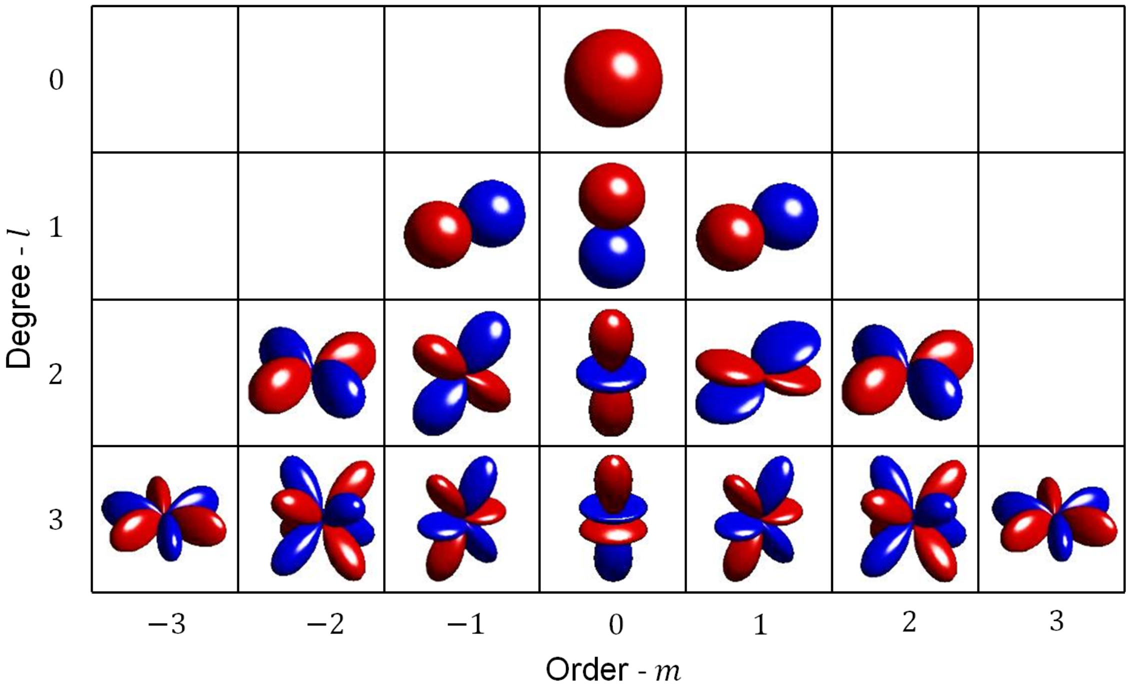


Figure A.1: Real part of spherical harmonic base functions up to the degree $l = 3$. For each direction away from the origin (θ, φ), the absolute value of the function's real part, $|\Re\{Y_l^m\}|$, is depicted by the radius and the sign by the color: red for positive values and blue for negative ones.

B

Parzen-Windows

Parzen-windows is a method for probability density estimation, originally introduced by Emanuel Parzen in 1962 [71]. The key concept is very simple: given a collection of independent and identically distributed samples of a random variable, several normalized window functions, also known as kernel functions, are superposed, each one centered on a different sample as illustrated in Figure B.1. More precisely, given a set of K d-dimensional samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\}$, the pdf estimate is given by:

$$\hat{p}(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} \phi(\mathbf{x} - \mathbf{x}_k, h), \tag{B.1}$$

where $\phi(\mathbf{x}, h)$ is the kernel function and h the window width. The kernel should be a finite-valued non-negative density function such that:

$$\int_{\mathbb{R}^d} \phi(\mathbf{x}', h)\mathbf{x}' = 1, \tag{B.2}$$

and the window width should be non-negative.

The kernel used in this work was the Gaussian kernel which is defined as follows:

$$\phi(\mathbf{x} - \mathbf{x}_k, h) = \frac{1}{(2\pi)^{d/2} h^d} exp\left(\frac{(\mathbf{x} - \boldsymbol{x}_k)^T (\boldsymbol{x} - \boldsymbol{x}_k)}{2h^2} \right). \tag{B.3}$$

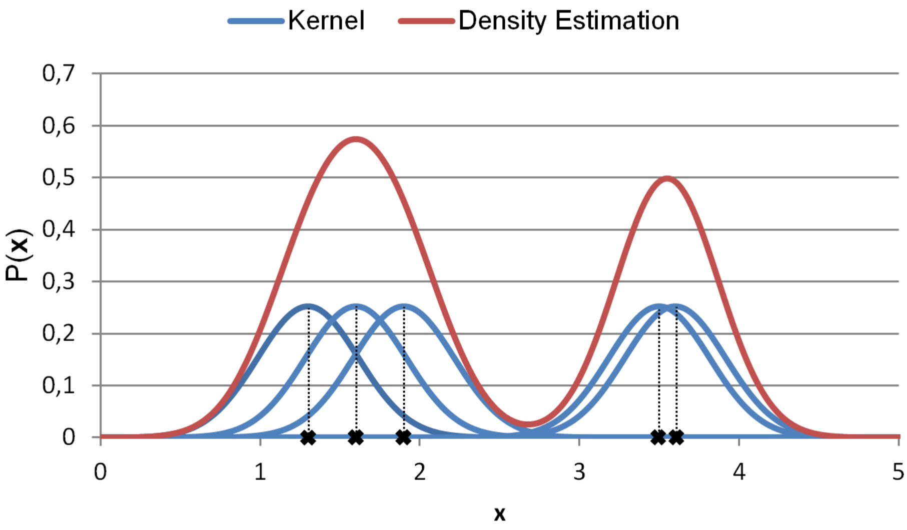


Figure B.1: Parzen-Window density estimation. The probability density estimation (red) is computed by the sum of several kernel functions (blue), each one centered at one sample (black crosses). The estimation of the density function still needs to be normalized.