UNIVERSIDAD
DE GRANADA

TESIS DOCTORAL

STATISTICAL NEUROIMAGE MODELING, PROCESSING AND SYNTHESIS BASED ON TEXTURE AND COMPONENT ANALYSIS: TACKLING THE SMALL SAMPLE SIZE PROBLEM.

Presented by:
Francisco Jesús Martínez Murcia

Advisors:
Juan Manuel Górriz Sáez
Javier Ramírez Pérez de Inestrosa

To apply for the:
International PhD Degree in Information and Communication Technologies.

Junio 2017

Francisco Jesus Martinez Murcia

Statistical Neuroimage Modeling, Processing and Synthesis based on Texture and Component Analysis: Tackling the Small Sample Size Problem.

Titleback

This document was written with LaTeX on Linux using a modified ArsClassica, a reworking of the ClassicThesis style designed by André Miede, inspired to the masterpiece *The Elements of Typographic Style* by Robert Bringhurst.

Contacts

☞ fjesusmartinez@ugr.es

ABSTRACT

The rise of neuroimaging in the last years has provided physicians and radiologist with the ability to study the brain with unprecedented ease. This led to a new biological perspective in the study of neurodegenerative diseases, allowing the characterization of different anatomical and functional patterns associated with them. Computer Aided Diagnostic (CAD) systems use statistical techniques for preparing, processing and extracting information from neuroimaging data pursuing a major goal: optimize the process of analysis and diagnosis of neurodegenerative diseases and mental conditions.

With this thesis we focus on three different stages of the CAD pipeline: preprocessing, feature extraction and validation. For preprocessing, we have developed a method that target a relatively recent concern: the confounding effect of false positives due to differences in the acquisition at multiple sites. Our method can effectively merge datasets while reducing the acquisition site effects. Regarding feature extraction, we have studied decomposition algorithms (independent component analysis, factor analysis), texture features and a complete framework called Spherical Brain Mapping, that reduces the 3-dimensional brain images to two-dimensional statistical maps. This allowed us to improve the performance of automatic systems for detecting Alzheimer's and Parkinson's diseases. Finally, we developed a brain simulation technique that can be used to validate new functional datasets as well as for educational purposes.

Guide:

https://plg.uwaterloo.ca/~migod/research/beckOOPSLA.html

RESUMEN

Resumen de la tesis en español.

DECLARACIÓN

D. Juan Manuel Górriz Sáez, Doctor por la Universidad de Cádiz y la Universidad de Granada y Catedrático del Departamento de Teoría de la Señal, Telemática y Comunicaciones de la Universidad de Granada y

D. Javier Ramírez Pérez de Inestrosa, Doctor por la Universidad de Granada y Catedrático del Departamento de Teoría de la Señal, Telemática y Comunicaciones de la Universidad de Granada,

MANIFIESTAN:

Que la presente Memoria titulada "Statistical Neuroimage Modeling, Processing and Synthesis based on Texture and Component Analysis: Tackling the Small Sample Size Problem. ", presentada por Francisco Jesús Martínez Murcia para optar al grado de Doctor por la Universidad de Granada, ha sido realizada bajo nuestra dirección. Con esta fecha, autorizamos la presentación de la misma.

Granada, Junio 2017

Fdo: Juan Manuel Górriz Sáez

Fdo: Javier Ramírez Pérez de Inestrosa

Memoria presentada por Francisco Jesús Martínez Murcia para optar al Grado de Doctor por la Universidad de Granada.

Francisco Jesús Martínez Murcia

DECLARACIÓN

El doctorando Francisco Jesús Martínez Murcia y los directores de la tesis Juan Manuel Górriz Sáez y Javier Ramírez Pérez de Inestrosa garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, Junio 2017

Directores de la Tesis:

Juan Manuel Górriz Sáez

Javier Ramírez Pérez de Inestrosa

Doctorando:

Francisco Jesús Martínez Murcia

ACKNOWLEDGEMENTS

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity,
and especially because it produces objects of beauty.

— Donald Ervin Knuth

I wish first of all to thank the members of the Italian TEX and LATEX User Group, in particular Claudio Beccari, Marco Brunero, Fabiano Busdraghi, Gustavo Cevolani, Rosaria D'Addazio, Massimiliano Dominici, Gloria Faccanoni, Daniele Ferone, Tommaso Gordini, Gianluca Gorni, Enrico Gregorio, Maurizio Himmelmann, Jerónimo Leal, Paride Legovini, Lapo Filippo Mori, Andrea Tonelli, Ivan Valbusa, Emiliano Giovanni Vavassori and Emanuele Vicentini, for their invaluable aid during the writing of this work, the detailed explanations, the patience and the precision in the suggestions, the supplied solutions, the competence and the kindness: thank you, guys! Thanks also to all the people who have discussed with me on the forum of the Group, prodigal of precious observations and good advices. Finally, thanks to André Miede, for his wonderful ClassicThesis style, and to Daniel Gottschlag, who gave to me the hint for this original reworking.

♡

Gracias también a todos los participantes en la encuesta del mapa de color.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

PCA	Principal Component Analysis
ICA	Independent Component Analysis
FA	Factor Analysis
SPECT	Single Photon Emission Computed Tomography
CT	Computed Tomography
PET	Positron Emission Tomography
AD	Alzheimer's Disease
PD	Parkinson's Disease
PKS	Parkinsonism

ASD	Autism Spectrum Disorder
MRI	Magnetic Resonance Imaging
fMRI	functional MRI
PLS	Partial Least Squares
SWPCA	Significance Weighted Principal Component Analysis
SBM	Spherical Brain Mapping
VBM	Voxel Based Morphometry
SPM	Statistical Parametric Mapping
SPM8	Statistical Parametric Mapping Software, version 8
CTL	Control Subject
VAF	Voxels As Features
CAD	Computer Aided Diagnosis
ADNI	Alzheimer's Disease Neuroimaging Initiative
PPMI	Parkinson's Progression Markers Initiative
VDLN	Virgen de las Nieves Hospital
VDLV	Virgen de la Victoria Hospital
MRC-AIMS	Medical Research Council Autism Imaging Multicentre Study
MNI	Montreal Neurological Institute
synT1	simulated T1 - weighted Inversion Recovery
qT1	quantitative T1 - weighted
qT2	quantitative T2 - weighted
GM	grey matter
GLM	General Linear Model
WM	white matter
CSF	cerebro-spinal fluid

ANOVA	Analysis Of Variance
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVC	Support Vector Classifier
CBM	Component Based Morphometry
KDE	Kernel Density Estimation
MCI	Mild Cognitive Impairment
EM	Expectation-Maximization
PDF	Probability Density Function
FWE	Family Wise Error rate
RF	radiofrequency
SNR	Signal-To-Noise Ratio
rCBF	regional Cerebral Blood Flow
DAT	Dopamine Transporters
FBP	Filtered Back Projection
FDR	False Discovery Rate
ROI	Region of Interest
HMM	Hidden Markov Model
AC	Anterior Commissure
LBP	Local Binary Patterns
CV	Cross-validation
LOO	Leave-One-Out
TP	True Positive
FP	False Positive
TN	True Negative

FN	False Negative
KL	Kullback-Leibler
MWW	Mann-Whitney-Wilcoxon
EVD	eigen-value decomposition
SWEDD	subjects without evidence of dopaminergic deficit
GLCM	Grey Level Co-occurrence Matrix

Part I

INTRODUCTION

1 INTRODUCTION

1.1 Motivation

In recent years, there has been a rise in the use of neuroimaging in the clinical practice. It has improved and speeded the procedure of diagnostic, providing unprecedented insight into the brain. Neuroimaging is very extended in research as well. Different fields such as psychiatry, neurology, psychology, behavioural science or biology make extensive use of brain imaging in their studies.

The basis of these studies are common: a selection procedure by which a representative set of subjects is recruited, the fulfilment of an experiment on (or by) each subject and a statistical analysis of the acquired data. Particularly, when studying a certain disease, it is common to recruit subjects affected by the disease and non-affected, healthy subjects, usually known as Control Subjects (CTLs). Then, in this typical example, both affected and CTLs are scanned, and brain anatomy or function is analysed using statistical tools. The result of this analysis is a list of significant differences between structure or function that could be linked to the disease.

Computer Aided Diagnosis (CAD) systems provide a set of tools to help setting up and performing these studies. It is currently a thriving area of research involving multidisciplinary teams, combining computer science, mathematics, medicine, artificial intelligence, statistics, machine learning, and many others [**Martinez-Murcia2016**]. The main aim is to assist clinicians in the procedure of diagnosis and study of the diseases by providing software that can effectively recognize disease patterns, characterize differences and make predictions.

One fundamental issue often found in this studies is the sample size. The number of subjects frequently ranges from tens to hundreds, whereas the number of features (namely voxels) to be analysed can add up to millions. This causes the so-called *Small Sample Size Problem* [**Duin2000**] which negatively affects the statistical power of any experiment performed using these datasets [**Button2013**].

Figure 1.1: Illustration of the separation between points in one-dimensional and two-dimensional spaces.

1.2 The Small Sample Size Problem

The *Small Sample Size Problem* refers to a problem that arises when the proportion between number of subjects and number of features is large. Think of, e. g., 15 points in a one-dimensional line, as in Figure 1.1. If we think of a subject as a vector, we would have 15 subjects in a one-dimensional space. Now, imagine that we add a second dimension. It is easy to see that our subjects would be farther than in the two-dimensional world. And the same would happen if we move to four, ten or thousands of dimensions. The farther our points are, the more difficult is for a statistical tool to extract information. That is what we call *almost empty spaces* [**Duin2000**, **Stoeckel04**], in contrast to *dense spaces*, where points are closer.

Neuroimaging provides hundreds of thousands, or even millions of voxels, in what could mean millions of features. That implies that any calculation performed in those almost empty spaces will eventually lack information. This implies a loss of statistical power of the methods used, usually producing false negatives (the system is unable to detect real signal) and false positives (the system detects signal where there is not). These are known in statistics as Type I and Type II errors respectively.

In differential diagnosis studies, the small sample size problem leads to wrong conclusions about where real differences are located. This, in addition to untracked confounding variables are one of the fundamental sources of non-reproducibility un current neuroimaging studies [**Button2013**].

The solution might seem straightforward: increase sample size. But this is not always possible, since neuroimaging studies do their best at recruiting as many

people as they can with a limited budget. Many efforts have been put into establishing multi-centre collaborations that allow the recruitment of a larger population, but despite offering a higher statistical power, these studies still suffer from a number of confounding variables such as population bias or scanner differences [**haar2014anatomical**]. In Chapters 7 and 8 we explore different approaches to this solution.

Another option involves reducing the number of features, via feature selection or feature extraction. This has been widely used in computed-aided methodology for neuroimaging [**DeMartino2007**, **xu2009source**, **Gorriz2010**, **Illan2011**, **Martinez-Murcia2016**] with great success, and solutions using this approach will be treated in Chapters 4, 5 and 6.

The Small Sample Size problem is directly related to the *Curse of Dimensionality* [**Krishnaiah1982**], which proves that, in contrast to what could be expected, once a certain classifier performance has been achieved, it holds or even decreases when feeding more features to the classifier. The problem also affects statistical hypothesis testing, a tool widely used for inference in neuroimaging, in what is known as the *Multiple Comparisons* problem [**Benjamini2010**], a particular field that is still being studied.

1.3 Aims and Objectives

This thesis aims to contribute new approaches to overcome the small sample size problem in neuroimaging. This can provide more accurate CAD systems by reducing the number of false positives, increasing the reliability of their results.

We will take two different approaches, as commented in previous sections: increasing the sample size and reducing the feature space. Therefore, we can define the following objectives:

- Develop and evaluate algorithms that reduce the feature space, in which is usually known in the field as feature extraction and feature selection strategies.
- Develop and evaluate new strategies to increase the sample size in neuroimaging studies.

Most of the studies in the literature focus on the first objective. Feature extraction algorithms that use Principal Component Analysis (PCA) [**Khedher2015**, **Towey2011**] or Partial Least Squares [**Segovia2013**], among others, are widely studied. We have developed three different approaches to those:

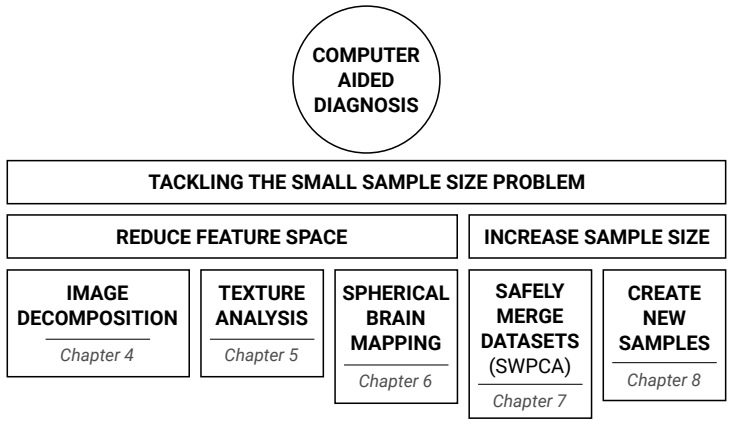


Figure 1.2: Structure and connexions between the different strategies proposed in this thesis, organized by chapters and parts.

- A combination of image decomposition algorithms and feature selection. In this approach we have used three criteria to select the most significant voxels from the images, and then applied Factor Analysis (FA) and Independent Component Analysis (ICA) to decompose the data and significantly reduce their feature space.

- A feature extraction based on texture analysis.

- A novel strategy called Spherical Brain Mapping (SBM). This feature extraction technique uses a spherical coordinate system to map statistical measures to a bidimensional plane. It builds paths used as feature selection vectors where several measures are computed.

On the other hand, we have evaluate newer ways to increase sample size in neuroimaging studies. We have developed:

- A system to reduce undesired variance in structural multicentre studies, called Significance Weighted Principal Component Analysis (SWPCA). This system is intended to reduce the amount of false positives in large collaborations, providing more homogeneous images and improving their statistical analysis.

- An algorithm to simulate functional brain images using existing data, and therefore, increase sample size.

1.4 Organization of this Thesis

This thesis work is organized in four parts plus appendices, each of which is subdivided in several chapters. In the first part, we introduce the motivations and main aims of this work (Chapter 1), examine the state of the art in medicine, neuroimaging and CAD systems (Chapter 2) and present a general methodology that will be followed throughout this thesis, including preprocessing and evaluation (Chapter 3).

Parts ii and iii refers to each of the solutions outlined in the previous section, and disaggregated in Figure 1.2. In part ii we focus on the feature reduction techniques, including decomposition methods (Chapter 4), texture analysis (Chapter 5), and the novel algorithm Spherical Brain Mapping (Chapter 6). On the other hand, Part iii is focused on two different strategies used to increase the sample size: the Significance Weighted Principal Component Analysis algorithm (Chapter 7), used to safely merge structural images acquired at different centres, and a neuroimage simulation algorithm (Chapter 8) that can be used to extend existing functional datasets.

Finally, in Part iv we provide a general discussion of the results presented in this thesis, conclusions about the methods and prospective work that could be performed with this basis.

1.5 Contributions

Some ideas and figures have appeared previously in the following publications, that we divide here in articles and conference presentations.

1.5.1 Articles

1.5.2 Conferences

1.5.3 Books

2 STATE OF THE ART

We have already stated the motivation and objectives of this thesis. Now, we will describe in detail some of the more relevant issues for the state of the art. First, in Section 2.1, we will make an introduction to the different neuroimaging modalities used in our experiments. Afterwards, we will provide some insights into the neurological and psychiatric disorders treated here in Section 2.2 or the most extended voxel-wise analyses used in the neuroimaging community at Section 2.3. Finally, at Section 2.4 we will explore recent contributions to the field that use Machine Learning.

2.1 Introduction to Neuroimaging

Medical imaging refers to all types of 2D, 3D and 4D images used in clinical practice. These involve many different modalities, among them X-rays, ultrasound, endoscopy, microscopy, etc. In neuroimaging, the most extended is by far Magnetic Resonance Imaging (MRI), which provides intensity maps that represent the internal structure of the brain. Other modalities are aimed at studying the function of the brain, by injecting radioactive ligands that, linked to a receptor, can measure its distribution. This is the case of Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT).

2.1.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is perhaps the most widespread in neuroimaging, given its ability to visualize both structural and functional (in functional MRI) properties of the brain, and, in contrast to other imaging modalities, is considered non-invasive. MRI uses strong magnetic fields to excite certain atomic nuclei, that can absorb and emit this energy.

MRI combines the magnetic field with a radiofrequency (RF) emission to excite the atomic nuclei present in corporal structure, resulting in a image of the distribution of certain atoms in the body. Most MRI use hydrogen atoms, since they are present in water (which adds up to around 70% of body mass) and

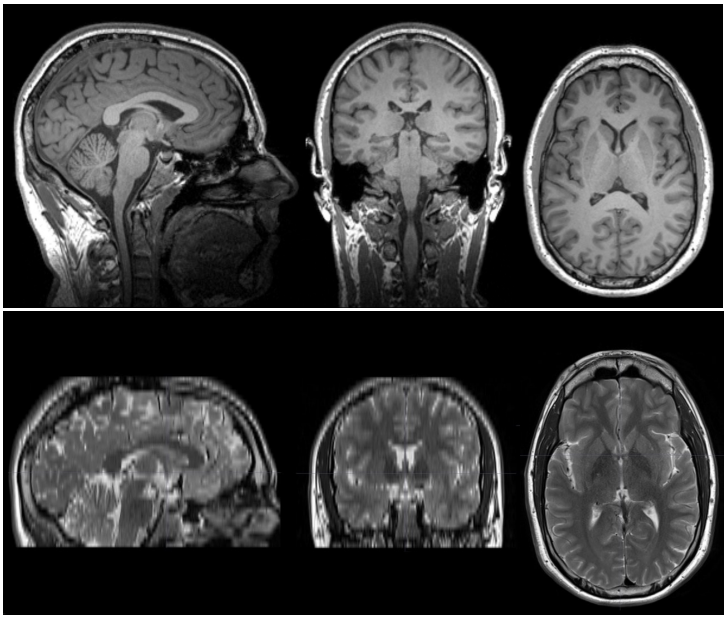


Figure 2.1: Example of T1 and T2-weighted MRI images of the same subject (me).

the signal derived is stronger than other atoms, increasing the Signal-To-Noise Ratio (SNR), and therefore, the image quality.

The procedure uses a strong magnetic field B_0 to align the magnetic moment of the hydrogen nuclei in parallel or anti-parallel (depending of their initial spin). This way, the magnetic moment of all nuclei will increase up to a stable state, in contrast to their null value in absence of B_0. Within this magnetic field, the hydrogen atoms precess around an axis along the direction of the field.

A given nuclei has a resonance frequency which is proportional to the intensity of B_0, which, by using strong fields, allow us to resonate hydrogen far below potentially damaging frequencies. The precession frequency is determined by the Larmor equation (2.1):

$$f_0 = \frac{\gamma}{2\pi} B_0 \tag{2.1}$$

where γ depends on the nuclei, which in the case of hydrogen, $\gamma = 42.6$ MHz/T. When a subject is introduced in the MRI scanner, it is submitted to the magnetic field B_0, so that the hydrogen nuclei are aligned to the field, with a precession frequency f_0. Then, a RF pulse of the same frequency is generated, which is then absorbed by the nuclei, forcing them to place perpendicular to the field. Once the RF emission is interrupted, the nuclei return to its equilibrium state by means of a procedure called relaxation. In this procedure, they emit part of

the absorbed energy, which is then captured by a RF receptor. Usually, position information is encoded in the RF signal by varying B_0 using gradient coils.

The RF signal is measured during the relaxation time, and two different relaxation times are set: the T1 (spin-lattice) relaxation time and the T2 (spin-spin) relaxation time. The T1 time is the time during which nuclei emit energy to the adjacent tissue and realign to the longitudinal plane (z axis), whereas the T2 time refers to the time when nuclei realign to the transversal plane (y axis). These times are used to create T1-weighted and T2-weighted images (see Figure 2.1). T1-weighted images allow to distinguish between GM and WM in the cerebral cortex, to identify fatty tissue, and generally, obtain structural information. Conversely, T2-weighed images are used to assess cerebro-spinal fluid (CSF) or to visualize and identify WM lessions.

2.1.2 Single Photon Emission Computed Tomography

The Single Photon Emission Computed Tomography (SPECT) is based on the principles of Computed Tomography (CT), by which a series of signal acquisition at different angles can be reconstructed back into a bidimensional distribution of the signal. In SPECT, a gamma photon emitting radioisotope is linked to a pharmaceutical that binds to a given biomarker, generating a radiopharmaecutical or agent. This agent is injected into the patient, and after a certain time in which the radiopharmaceutical is distributed, the patient is introduced into the SPECT-CT scan.

Afterwards, the scanner performs a series of acquisitions at different planes and angles from the body, from which the gamma signal is measured. For each plane, all acquisitions at each angle are pooled and a single two-dimensional image is reconstructed using a Filtered Back Projection (FBP) algorithm, or Radon inversion formula [**Herman2009l**], which derives from the Fourier's Theorem. A total of 180 projections per plane, using an angular resolution of 2 degrees, are usually taken.

There exist a number of radiopharmaceutical used in clinical practice, and therefore, we will focus on the two varieties used in this thesis. First, we use an agent called ^{99m}Tc-HMPAO, which consists of two stereoisomers of hexametazime (HMPAO) linked to the radioisotope technetium 99-metastable. This agent is usually used to assess regional Cerebral Blood Flow (rCBF), which can be used to diagnose neurological diseases or cancer.

Additionally, we use images generated using the agent Ioflupane (^{123}I), a cocaine analog with high binding affinity for Dopamine Transporters (DAT). It is used fundamentally in the assessment of Parkinson's Disease (PD), given that

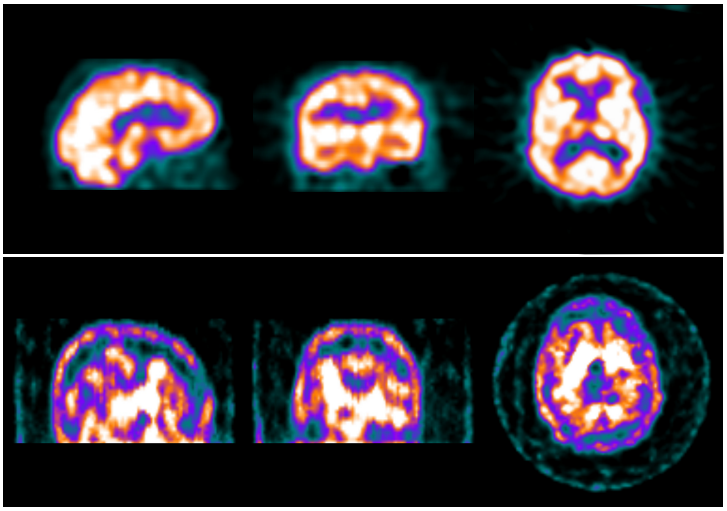


Figure 2.2: Example of SPECT images, a SPECT-HMPAO and a SPECT-DaTSCAN.

the disease is associated with a loss of dopaminergic neurons in the striatal region.

2.1.3 Positron Emission Tomography

The Positron Emission Tomography (PET) is a technique similar to SPECT, but in this case, the agent used and the equipment is designed to deal with a pair of gamma photons resulting of the annihilation of a positron with its corresponding antiparticle, the electron. The pair of photons are generated in opposite directions, and the detection depends on them being simultaneously or coincidently detected at the receptor. The receptor comprises a scintillator which emits light when the gamma photon incides, and a detector, usually a photomultiplier tube or silicon avalanche photodiodes.

It uses the same FBP algorithm as SPECT in the reconstruction of the images, and a similar strategy for acquiring the signal at different angles. However, the amount of data is smaller than in SPECT, and therefore, the reconstruction procedure is harder. As a result, PET scanner operation is considered more costly than SPECT **[Carlson2016]**.

The agent used in the images that we have processed is PET-FDG, also known as Fludeoxyglucose (^{18}F). It is a glucose analogue that allows us to measure the glucose metabolism in the brain. It is widely used in neurology **[Newberg2002]** and cancer detection **[Kelloff2005]**, since it can be correlated with cellular activity.

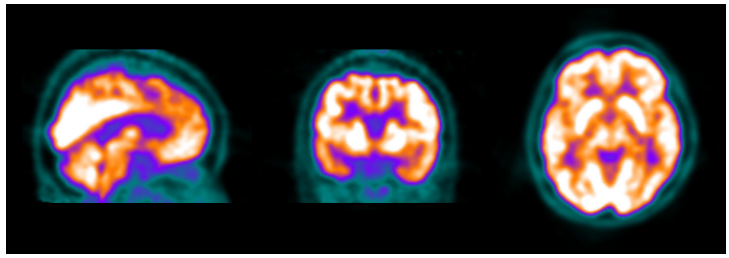


Figure 2.3: Example of a PET-FDG image.

2.2 Medical Background

2.2.1 Alzheimer's Disease

Alzheimer's Disease (AD) is the most common cause of dementia in the world, with more than 46 million people affected, and it is likely to increase up to 135.5 million by 2050 [**Association2016**]. Its causes are still not clear, but it is characterized by deposits of high amounts of structures such as Amyloid-β (Aβ) plaques or neurofibrillary tangles accompanied by synaptic dysfunction and neurodegeneration that eventually lead to cell death [**Ballard2011**, **Sevigny2016**].

Diagnosis of AD is often based on the clinical history of the patient, using cognitive tests along with medical imaging and blood tests. A definite diagnosis can only be addressed post-mortem, via a direct examination of the brain tissue [**Ballard2011**]. Cognitive tests such as *Mini Mental State Exam* or *Clinical Dementia Ratio* are widely used in clinical practice.

Initial symptoms of AD are often mistaken for normal ageing, leading to a state known as Mild Cognitive Impairment (MCI). Not all MCI-affected subjects develop AD. In fact, the prediction of MCI conversion is the most urging challenge in AD research, since an early diagnosis can lead to an improvement in life expectancy and quality of life of the patients.

MRI brain images have been extensively used in the diagnosis of AD by assessing neurodegeneration on GM and WM tissues. Research has shown in [**Baron2001**, **Misra2009**, **Pievani2013**, **Dubois2007**] that neurodegeneration in Alzheimer's Disease mainly occurs in the GM tissue. Particularly grey matter loss has been described in the Hippocampus and Parahippocampal lobes, according to the NINCDS-ADRDA criteria for AD diagnosis [**Dubois2007**], with further atrophy described in the medial temporal structures, the Posterior Cingulate gyrus and adjacent Precuneus [**Baron2001**]. Moreover, significantly lower volumes of certain regions in GM and WM have been considered a promising biomarker and predictor of the progression of AD in a longitudinal study involv-

ing MCI patients [**Misra2009**], and some structures in the striatum (putamen and caudate nucleus) have shown important volume abnormalities [**Pievani2013**]. All these data suggest that many of the symptoms of AD can be observed in anatomical MRI images even in early stages of the disease, which could be of great help in its successful diagnosis and treatment.

Nuclear imaging, such as PET or SPECT, have also been used in clinical practice, especially to discard other diseases. In typical PET-FDG or SPECT-HMPAO, AD is characterized by reduced brain activity in bilateral regions, such as the posterior cingulate gyri and precunei, as well as the temporo-parietal region [**Claus1994**]. It also affects the frontal cortex and the whole brain in severe cases [**Leon1983**, **Kogure2000**].

Recently, new more specific radiopharmaceuticals have been developed, among them the Pittsburgh compound B (PiB). This drug binds to fibrillar Aβ allowing *in vivo* visualization of Aβ plaques in the brain [**Ikonomovic2008**]. However, due to their technical requirements -a relatively small half-life of the radioactive element-, they are unusual in clinical practice.

2.2.2 Parkinsonism

Parkinsonism (PKS) or Parkinsonian Syndrome is the second most common neurodegenerative disease in the world, with a prevalence of 1-3% in the elder population (over 65 years)[**Eckert2007**]. It is characterized by hypokinesia, rigidity, tremor and postural instability [**Eckert2007**]. It is not a single disease itself, but a wide range of conditions that share similar symptoms. The most common cause of PKS is Parkinson's Disease (PD), but other possible causes include toxins, a few metabolic diseases, and other extrapyramidal syndromes such as Multiple System Atrophy, Progressive Supranuclear Palsy or Cortico-Basal Degeneration [**Christine2004**, **tatsch2008extrapyramidal**].

2.2.2.1 *Parkinson's Disease*

The etiology of Parkinson's Disease (PD) involves the progressive loss of Dopamine Transporters (DAT) of the nigrostriatal pathway. This causes a decrease in the dopamine content of the striatum, since the pathway connects the substantia nigra to the striatum.

In PD, structural imaging such as MRI has limited value, since structural abnormalities can be seen only in the latter stages. Nevertheless, in molecular imaging, a number of radiopharmaceuticals have been proposed to assess the levels of pre and post-synaptic DAT at the striatum. Among them, the ^{123}I-ioflupane (better known by its tradename DaTSCAN) is perhaps the

most popular. DaTSCAN binds to the DAT at the striatum [**Winogrodzka2003**, **PunalRiob002007**, **Eckert2007**], allowing the estimation of DAT density by means of a SPECT scanner. In DaTSCAN images, a reduced uptake at the striatum is a clear indication of DAT loss, and therefore, of PD [**PunalRiob002007**].

2.2.2.2 *Extrapyramidal Syndroms*

Among the extrapyramidal syndroms of PKS, the most relevant diagnoses are Multiple System Atrophy, Progressive Supranuclear Palsy and Cortico-Basal Degeneration [**tatsch2008extrapyramidal**]. An accurate diagnosis can positively impact in the health of these patients, avoiding wrong treatment decisions.

Structural imaging, as in the previous case, has little value here. To establish a differential diagnosis with PD, different drugs have been proposed. DaTSCAN is widely used to assess pre-synaptic dopaminergic loss, but in a post-synaptic level, one of the most extended is ^{18}F-DMFP-PET [**Segovia2016a**]. However, in this thesis we have focused only on the pre-synaptic level, and therefore, in the diagnosis of PD.

2.2.3 Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is a neurodevelopmental syndrome characterized by social and communication impairment as well as restricted, repetitive patterns of behaviour, interests or activities. Its origins are still unknown, although research suggest [**Szatmari1999**] that there exist genetic risk factors.

The evidence of either functionally or structurally affected areas in the brain is a major concern [**Ecker2014**, **haar2014anatomical**]. In the latter years, many strategies have been explored to recruit large samples in order to detect significant differences between affected and non-affected subjects. This has been addressed by initiatives such as the Medical Research Council Autism Imaging Multicentre Study (MRC-AIMS) [**Ecker2012**, **Ecker2013**] or the Autism Brain Imaging Data Exchange (ABIDE) [**DiMartino2014**].

2.3 Voxelwise Analyses

Traditional analysis of neuroimaging involves visual analysis by experts clinicians, or semi-quantitative analysis of Regions of Interest (ROIs). With the rise of neuroimaging in the mid-nineties, some computer-aided solutions appeared, of which the most extended are the widely known Statistical Parametric Map-

ping (SPM) [**Friston1994**] and its extension to structural imaging Voxel Based Morphometry (VBM) [**Ashburner2000**].

2.3.1 Statistical Parametric Mapping

Statistical Parametric Mapping (SPM) is a new methodology to automatically examine differences in brain activity in functional imaging studies involving functional MRI (fMRI) or PET, firstly proposed by Friston in [**Friston1994**]. The technique can be applied either to static images (e.g., PET) or timeseries (fMRI), using inference techniques based on hypothesis testing, in order to construct the General Linear Model (GLM) that better describes the variability in the data.

Statistical hypothesis testing involves constructing a pair of hypotheses: H_0, or the null hypothesis, that states no relationship between variables; and H_1, the alternative hypothesis. In neuroimaging, H_0 usually means that there are no relevant differences between classes (for example, between patients affected by Alzheimer's Disease (AD) and CTL), and H_1 implies that there is a significant difference. Many different tests such as massive univariate t-Test or Analysis Of Variance (ANOVA) (see Chapter 4 for more information on these techniques) can be used in the SPM software [**spm_book**], by using a design matrix that describes a t or F based contrast (for t-Test and ANOVA respectively). These terms are generally referred to as Z-values, namely the signed number of standard deviations an observation is above the mean.

The test are computed voxel-wise, from which a p-value can be obtained, nominally the probability of obtaining equally or more extreme Z values that the one actually found. p-values are very extended in neuroimaging, representing the probability of a Z value being equal or more extreme than the reference value given. In many studies $p < 0.05$ is used for measuring statistical significance, which means that only a 5% of the times a experiment is repeated we would obtain that result or a more extreme one. The use of the significance threshold $\alpha = 0.05$ implies that any voxel with a p-value smaller than 0.05 is considered sufficient to reject the null hypothesis.

SPM outputs maps like the one shown in Figure 2.4. There, significant Z-values according to a given threshold (FWE uncorrected or corrected, see Section 2.3.3) are displayed over an anatomical reference. The resulting maps allow a visual inspection of the active brain areas, which can later be related to a certain disease or task.

Although SPM's main feature is the estimation of differences, the term has been extended to cover the whole process performed by the SPM software. That is, it generally involves registration to a template, intensity normalization, smooth-

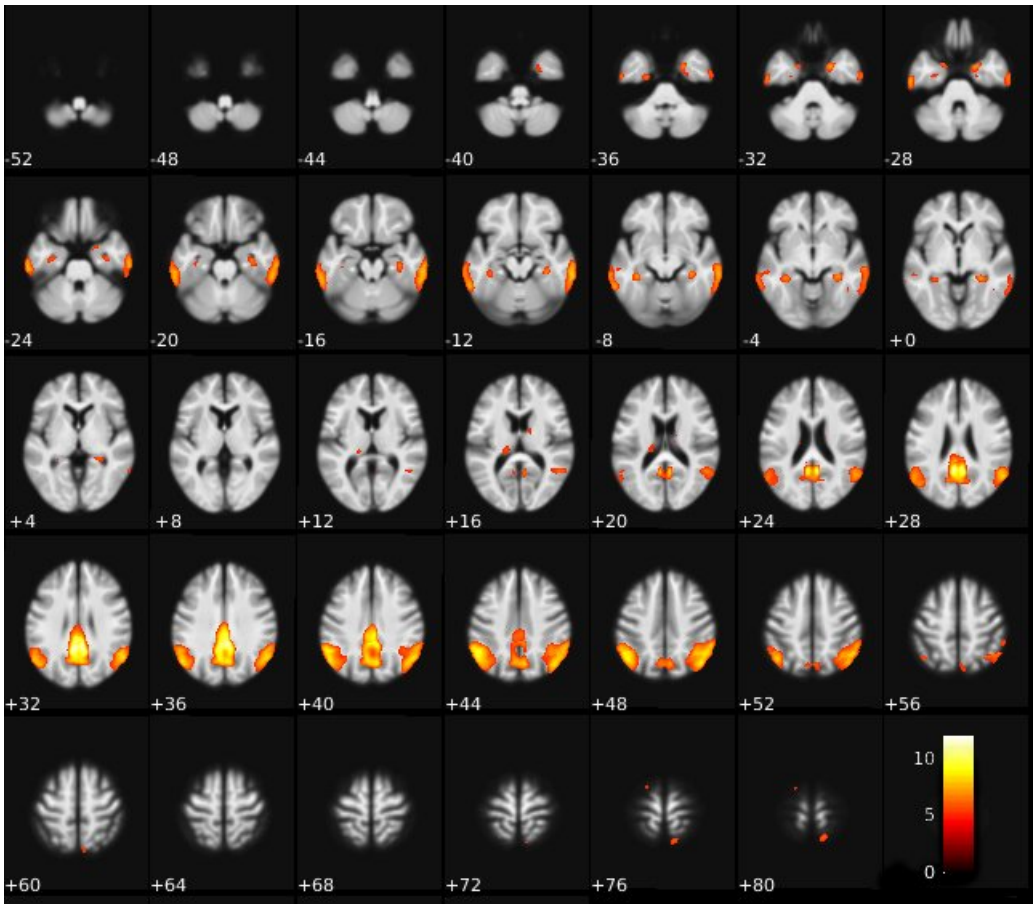


Figure 2.4: Example of a SPM analysis on a PET dataset displaying the differences between AD and CTL, using $p < 0.05$ and FWE correction.

ing, the proper SPM difference estimation and the display of the results. An overview of these procedures is provided at Chapter 3.

2.3.2 Voxel Based Morphometry

Voxel Based Morphometry (VBM) can be considered an extention of SPM applied to structural MRI images [**Ashburner2000**]. The procedure involves preprocessing (see Chapter 3), where smoothing is applied to reduce smaller anatomical differences. Afterwards, a GLM is applied to each voxel in the images, and a Z-score map similar to Figure 2.4 is produced.

Smoothing is more important in VBM than in regular SPM, since MRI images have higher resolution and are less noisy than functional images. Larger smoothing kernels will miss out smaller regions, while smaller kernel can lead to artifacts in the generated Z-maps, including misalignment of brain structures, differences in folding patterns or misclassification of tissue types [**Martinez-Murcia2016book**]. Therefore, the kernel size must be carefully chosen, usually using a-priori knowledge about the regions affected, and always double checking for artifacts and reproducibility.

The idea behind VBM has been extended in a number of papers, using multivariate approaches that takes into account all voxels at once, and not their individual differences. Some of them include ICA decomposition of the dataset and a posterior conversion to Z-scores in what was called Source Based Morphometry [**xu2009source**], or multidimensional Tensor Based Morphometry [**bossa2010tensor**].

2.3.3 The Multiple Comparisons Problem

The Multiple Comparisons problem arises when using hypothesis testing to assess statistical significance. This is widely used in neuroimaging, where statistical tests such as the t-Test or ANOVA are used to quantify voxel-wise differences, and state their statistical significance, or p-value. The p-value, as described above, is the probability of any value being more extreme than a certain threshold under a given hypothesis. In our problem, given the t-value T_i for the i^{th} voxel ($i = 1, \dots N$) of the images, and a threshold T_{th} under the hypothesis H, the significance can be assessed by checking:

$$P\left(T_i > T_{th} | H_0\right) < \alpha \tag{2.2}$$

where α is the significance level.

Choosing α is not trivial in neuroimaging. The use of the significance level $\alpha = 0.05$ implies that any voxel with a p-value smaller than 0.05 is considered

sufficient to reject the null hypothesis. This does not directly imply the necessity of accepting the alternative hypothesis H_1, although it is often thought so. Neither it yields the probability of the null hypothesis [**Dixon2003**].

If we apply $p < 0.05$ directly to a medical image of, for example, 300,000 voxels, that could mean the possibility of almost 15,000 voxels being false positives. Controlling the apparition of false positives when applying a massive univariate test is not trivial. It implies a balance between the true positive rate (sensitivity) or true negative rate (specificity), given that, for example, controlling the amount of false negatives will result in many false positives and vice-versa.

Usually, two options for controlling the amount of false positives are given: the Family Wise Error rate (FWE) and the False Discovery Rate (FDR). The FWE is the probability of obtaining at least one type I error. Mathematically, the null hypothesis for the i^{th} voxel H_{0i} states that there is no activation in that voxel. Therefore, the family-wise null hypothesis for our problem is:

$$H_0 = \bigcap_i H_{0i} \tag{2.3}$$

If we reject a single null hypothesis ($T_i > T_{th}$), we reject H_0. Therefore, we want to control the probability of a single voxel being significant if the family-wise null hypothesis is valid:

$$P\left(\bigcup_i \{T_i > T_{th}\} | H_0\right) < \alpha \tag{2.4}$$

In this case, we must obtain the critical value T_{th}, which is the higher t value that matches that expression. Many options have been proposed to this problem, among them the conservative Bonferroni correction, methods that use random field theory or permutation tests.

2.3.3.1 *The Bonferroni Correction*

The Bonferroni correction [**Shaffer1995**] rewrites eq. 2.2 setting $\alpha = \frac{\alpha}{N}$ so that:

$$P\left(T_i > T_{th} | H_0\right) < \frac{\alpha}{N} \tag{2.5}$$

That way, using the Boole's inequality:

$$FWE \leqslant \sum_i^N \frac{\alpha}{N} = \alpha \tag{2.6}$$

Therefore, we can comply with the imposed restriction for a maximum FWE, or in our case, a maximum rate α of false positives. This is considered a rather

conservative approach. In the example cited above, if we want to keep the FWE below 0.05, we should divide it by N, therefore obtaining a T_{th} that makes $\alpha = 0.05/N = 1.67 \times 10^{-7}$.

Other less conservative options try to compute a critical value T_{th} that minimizes the FWE using spatial information. This is the case of using an approximation of the distribution of the maximum statistic over the image, or the spatial correlation, including elements from random field theory (the approach used in SPM [**spm_book**]).

2.3.3.2 *Random Field Theory*

In the random field approach, the maps of the statistic are treated, under the null hypothesis, as a lattice representation of smooth isotropic three dimensional random fields of test statistics. This approximation to the problem allow us to approximate the upper tail of the maximum distribution, the part needed for defining an event that occurs when the map exceeds the critical value T_{th}. Further information about random field theory and how it is applied to neuroimaging can be found at [**spm_book**].

The other approach, based on the FDR, aims at controlling the proportion of false positives in the total number of voxels declared significant. The most extended procedure for controlling the FDR is that proposed by Benjamini and Hochberg [**Benjamini1995**]. The Benjamini and Hochberg method start with calculating the p-values of all voxels and ranking them so that:

$$p_1 \leqslant p_2 \leqslant \cdots \leqslant p_i \leqslant \cdots \leqslant p_N \quad \forall i = 1 \ldots N \tag{2.7}$$

2.3.3.3 *FDR Controlling Procedures*

Let q be the a maximum FDR value that we can afford, for example 0.05. For each i, we compute:

$$p_i \leqslant \frac{i}{N} q \tag{2.8}$$

The maximum i value that holds Eq. 2.8 is used as α, the significance level, and its corresponding statistical value (T_i in the case of a t-test) is used as the critical value. This test, under the family-wise null hypothesis H_0, is equivalent to controlling the FWE. However, FDR methods are less conservative than other approaches such as the Bonferroni or other FWE-based corrections, leading to a gain in statistical power.

2.3.3.4 *Permutation Tests*

An empirical way to obtain p-values without relying on any parametric assumption is permutation testing [**Anderson2001**, **Winkler2014**]. Permutation tests evaluate a statistic such as the F-statistic or the t-test using randomly target variables, in our case, the classes. The procedure is applied many times (up to 10,000), and for each permutation, only the maximum value of the computed statistic is considered. These values are used to build the null distribution, from which the family-wise corrected p-values are computed. Results obtained in permutation tests are comparable to those obtained using Random Field Theory [**Winkler2014**], and far less conservative than when applying the Bonferroni correction.

2.4 Machine Learning in Neuroimaging

Machine learning is a current trend in neuroimaging. It provides computers with the ability to learn from data, using a set of statistical and computational tools. Rather than being explicitly programmed for a certain task, machine learning systems are able to find relevant data, discover patterns and predict the outcome of the input data. Its application to medicine is often known as Computer Aided Diagnosis (CAD) [**Martinez-Murcia2016**].

There are two major branches of machine learning: supervised and unsupervised learning. Th former explores the patterns that lead to a certain outcome, whereas on the other hand, unsupervised learning explores the underlying structure of the data. Most of the CAD systems rely on supervised learning, since their intention is to discover patterns that can effectively predict a disease.

For simplicity, in this thesis, when talking about CAD, we will always refer to automatic CAD systems. That is, those that, once trained with previously known data, can predict the outcome of new, unseen data. A typical CAD system, like the one in Figure 2.5, consists of input data (in our case, neuroimaging), feature extraction, feature selection and a classification step. The most basic is the Voxels As Features (VAF) approach, in which all voxels are considered as features, and then used as input to the classifier [**Stoeckel04**]. However, many more advances can be made in this field by exploring different types of

2.4.1 Voxels as Features

Voxels As Features (VAF) [**Stoeckel04**] is an example of the simplest CAD system. It was originally proposed for evaluating and performing automatic diag-

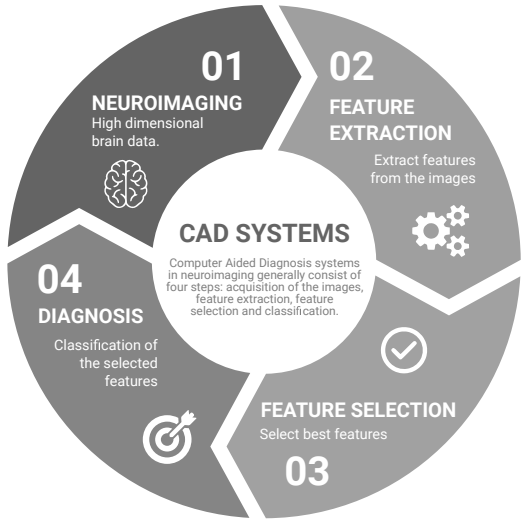


Figure 2.5: Illustration of a typical neuroimaging CAD system.

nosis of AD using functional SPECT imaging. It uses a standard preprocessing (registration, intensity normalization) and a Support Vector Classifier (SVC) (See Appendix B) to predict the class of an image using all its intensities as features. Feature extraction, here, considers all voxels, and then there is no feature selection applied.

It has been used in many works as a baseline [**Spetsieris2009**, **Salas-Gonzalez2009**, **Martinez-Murcia2016**], since it is comparable to the performance achieved by expert physicians using visual analysis [**Stoeckel04**]. The weight vector of the SVC can be inverse transformed to the dimension of the original images, and therefore provide a visual map that reflects the most influential voxels, in a similar way to the Z-maps of SPM and VBM.

2.4.2 Multivariate Analyses

Many improvements can be made to VAF by adding and refining feature extraction and feature selection techniques. With this addition, we can avoid the Small Sample Size, in addition to the ability to discover higher level abstractions that can be more representative of the progression of the studied diseases.

Feature extraction algorithms often change the strategy from a massive univariate approach, where a single feature is considered at each time, to multivariate analyses, where each feature can contain information from many voxels at the same time. Measures of total uptake of a given drug in nuclear imaging are a good example of this [**Zhou2007**, **Lozano2007**], but also the widespread

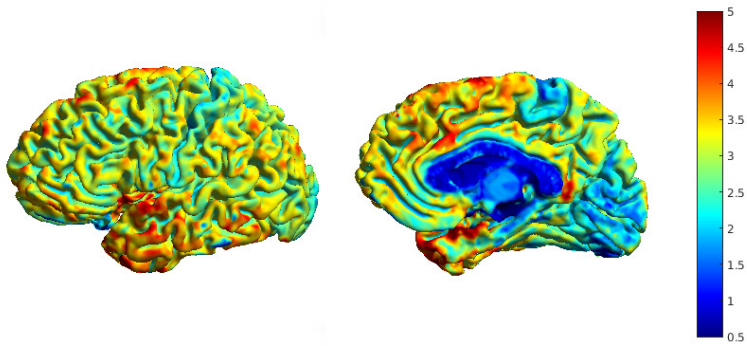


Figure 2.6: Example of the cortical thickness of a subject, obtained with the toolbox CAT12 in SPM12.

Cortical Thickness [**Dale1999**] provided by *FreeSurfer*. Cortical thickness is an estimation of the amount of GM in a direction perpendicular to its surface. It first estimates the GM-WM and the WM-CSF separation surfaces, and then characterizes the thickness of the tissue, allowing a characterization of GM differences such as atrophy or hypetrophy. Other pieces of software, such as SPM, also include toolboxes to compute cortical thickness, providing outputs such as the one that can be seen in Figure 2.6.

Other more advanced algorithms are image decomposition techniques such as PCA or Partial Least Squares (PLS), which have been extensively used in neuroimaging CAD systems [**Spetsieris2009**, **Illan2011**, **Towey2011**, **Segovia2013**, **Khedher2015**]. In these approaches, a given image can be represented as the linear combination of different components, and while the component loadings are common to all subjects, the weights of these components are unique to each patient. This allows us to identify the patterns that better discriminate between classes, leading to a more accurate diagnosis.

For its part, feature selection refers to different strategies aimed at finding an optimal subset of the extracted features, according to a certain criterion. Irrelevant features are therefore discarded, making our models faster and more cost-effective [**Guyon03**]. Feature selection algorithms are often subdivided in three approaches [**Martinez-Murcia2016b**]: filters, wrappers and embedded approaches.

Filters compute a feature relevance score from the data, which is then used to sort the different features. It is computed before the classification, and does not interact with it. Many scores can be derived from statistical features such as χ^2, t-Test, Fisher's Discriminant Ratio (FDR) or others [**Martinez-Murcia2013255**,

Martinez-Murcia2016b]. The output of these tests has been already used as a tool in voxelwise analyses, as we commented in Section 2.3.

Wrappers are similar to filtering methods, given that they assign a certain score to each feature. But in contrast to filters, the score is computed by estimating the performance in a predictive model, such as classifiers [**Kohavi1995**]. The most obvious measure here is accuracy, although other techniques such as Forward selection, backward elimination [**Guyon03**], genetic algorithms [**Kohavi1995**], or the expectation-maximization algorithm [**Gorriz2009**] have been used in the literature. And finally, embedded approaches use the very model that is being built to construct their optimal feature subset.

3 GENERAL METHODOLOGY

Throughout this thesis, we will propose many analysis techniques and CAD systems. We will apply them to many experiments, and use similar data and techniques in them. In this chapter, we will focus on the methodology that is common to most of these experiments, particularly focusing on preprocessing and evaluation of our systems.

To perform most automated analyses on neuroimaging, it is fundamental that images are comparable. Preprocessing comprises a series of algorithms that, applied after the acquisition and reconstruction of the images, produce directly comparable images in both structure and magnitude. Whether they have been used in one or all experiments, they can be classified in two major categories: spatial and intensity preprocessing. These are addressed in Section 3.1 and 3.2 respectively.

Afterwards, in Section 3.3, we will discuss how we evaluate our systems. Here we propose some performance measures and the procedure to obtain them by training and testing our systems.

3.1 Spatial Preprocessing

Spatial processing usually accounts for the differences in position, angles and structure that are commonly found between images. A common pipeline in, for example, MRI preprocessing, is the one found at Figure 3.1, where the images are registered (or spatially normalized) to a template, smoothed and finally segmented. The smoothing is an optional step, generally used in procedures like segmentation or VBM.

In this thesis, all the experiments in all image modalities involve spatial normalization. Smoothing, as well as segmentation, is only applied in some experiments that use MRI images, such as the segmented images in Chapter 6 or the whole-brain analysis performed in Chapter 7.

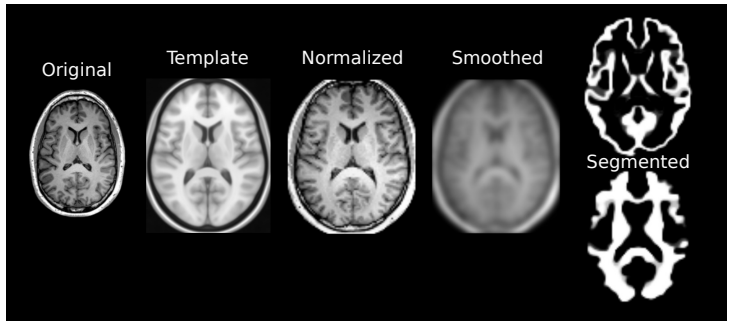


Figure 3.1: Typical pre-processing pipeline in MRI.

3.1.1 Spatial Normalization or Registration

Spatial Normalization, also known as Registration, is the procedure that by which every subject's brain is mapped from their individual space to a standard reference system. Registered images allows our system to overcome the individual differences in position and anatomy by establishing a common reference space in which a given coordinate represent the same anatomical position in all brains in the dataset.

There exist a number of pieces of software widely used for registering images, such as FreeSurfer [**Reuter2010**] or FSL (in the FLIRT and FNIRT package) [**Smith2004**], most of them perform linear, non-rigid and elastic transformations or a combination of these. In this work we have used the software SPM8 [**spm_book**] to perform registration of all the datasets, including MRI, SPECT and PET images. So, from this moment, we will focus on the registration as performed in the Statistical Parametric Mapping Software, version 8 (SPM8).

Linear registration usually refers to the affine transformation, a matrix multiplication that includes 12 parameters for translation, rotation, scale, squeeze, shear and others:

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.1}$$

This matrix multiplication is performed globally, as it transforms the whole image, not accounting for local geometric differences. In equations 3.2, 3.3 and 3.4 we give an example of the parameters that are computed for scale, translation and shear in 3D:

$$\text{scale} = \begin{bmatrix} s_x & 0 & 0 & 0 \\ 0 & s_y & 0 & 0 \\ 0 & 0 & s_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

$$\text{translation} = \begin{bmatrix} 1 & 0 & 0 & \Delta x \\ 0 & 1 & 0 & \Delta y \\ 0 & 0 & 1 & \Delta z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.3}$$

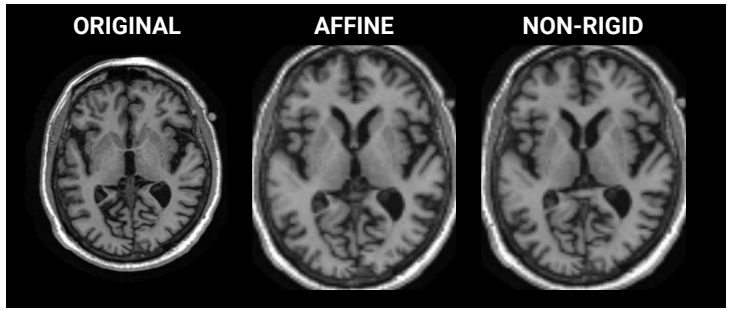
$$\text{shear} = \begin{bmatrix} 1 & h_{xy} & h_{xz} & 0 \\ h_{yx} & 1 & h_{yz} & 0 \\ h_{zx} & h_{zy} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.4}$$

The combination of all these operations result in the estimation of the twelve parameters that we found in Eq. 3.1, which are the ones used in SPM8. The estimation of these parameters is performed via the optimization of a cost function, that in SPM8 can be the minimum squared difference between the source image and the template **[spm_book]** in the case of within-modality registration, or the mutual information in between-modality registration. These functions are also used in FLIRT **[Jenkinson2001]**, whereas FreeSurfer uses the Tukey's biweight function (in `mri_robust_template`) **[Reuter2012]**.

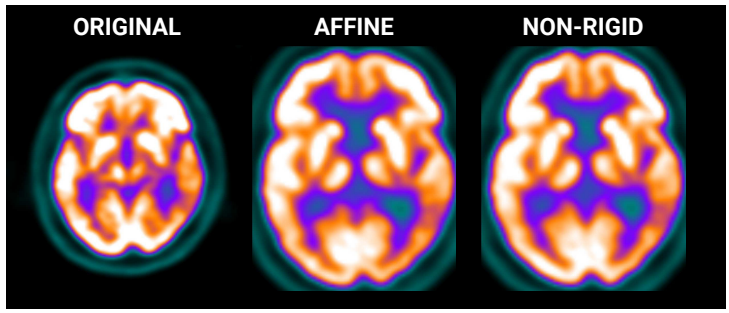
After the affine transform, the software usually performs a fine-tuning step via nonrigid transformations, to account for relevant anatomical differences between subjects. Nonrigid transformations range from the use of radial basis functions, physical continuum models and the large deformation models, or diffeomorphisms, that SPM8 uses. These procedures work by estimating a warp-field and then, apply it to the affine-registered images. An example of the differences of using only affine registration and applying diffeomorphisms can be found at Figure 3.2.

3.1.1.1 *Co-registration*

Sometimes we have several image modalities of the same subject, for example MRI and PET or functional MRI, often acquired at the same time. In this particular case, we can use the higher resolution MRI image to calculate the affine parameters and warping, and apply those to all modalities of the same subject. To do so, we perform a first co-registration, that is, a registration of the lower-resolution images (e.g. PET) to its correspondent MRI image. Being anatomically



(a) Comparison in MRI.



(b) Comparison in PET.

Figure 3.2: Comparison of the affine registration and the application of non-linear transformations to both MRI and PET images of the same ADNI subject.

similar, the co-registration usually comprises a single affine transformation. Afterwards, we can proceed with the registration of that MRI image to the template, and apply the same transformation to all its co-registered images.

3.1.1.2 *The MNI Space*

In this thesis, all images are coregistered to the Montreal Neurological Institute (MNI) space [**Mazziotta2001**]. This is the most widely used coordinate system, recently adopted by the International Consortium for Brain Mapping (ICBM) as its standard template. The three-dimensional coordinate system defined in MNI was intended to replace the Tailarach space, a system based on a dissected brain, that was used to compose an atlas by Tailarach and Tournoux [**Talairach1988c**]. The current template is known as ICBM152, and features the average of 152 normal MRI scans matched to an older MNI template using a nine parameter affine registration.

3.1.2 Segmentation

When using MRI images in this thesis, we often refer to grey matter (GM) and white matter (WM) maps, which is the result of the segmentation of the original data. Segmentation aims at producing maps of the distribution of different tissues, and it generally addresses GM, WM and CSF classes, although lately some software can output data for bone, soft tissue or very detailed functional regions and subregions [**Fischl2002**].

In this thesis we have used the VBM toolbox of the SPM8 software, which yields GM, WM and CSF maps. It features an Expectation-Maximization (EM) algorithm to model the distribution of the tissue classes as a mixture of gaussians and, by combining this distribution-based information with tissue probability maps using a bayesian rule, the software produces joint posterior probability maps for each tissue. To clean up the segmentation maps, a series of iterative dilations and erosions are used. Finally, since brain regions are expanded or contracted at the spatial normalization step, we can scale the segmented maps using modulation, producing final maps where the total amount of grey matter is preserved.

3.2 Intensity Normalization

Generally, structural modalities such as T1 and T2-weighted images are considered unitless, in contrast to functional imaging, in which each voxel's intensity represent the distribution of some biomarker, such as glucose metabolism, dopamine transporters, etc. These amounts are affected by many sources of variability that can affect the final values: contrast uptake, radiotracer decay time, metabolism, etc. Therefore, along with the previous spatial normalization, there is a need to normalize the intensities of the images, so that the amount they represent are comparable.

In the case of intensity normalization, the method acts as a linear transformation of the image, preserving fundamental information such as contrast between regions. This approximation estimates the new intensity values I' as:

$$I' = I/I_p \tag{3.5}$$

where I_p is a constant parameter that is unique for each image. After this division, the new intensities would be directly comparable. The technique used to compute the normalization parameter varies, ranging from the simplest normalization to the maximum [**Salas-Gonzalez2009**, **Martinez-Murcia20129676**] to complex methodologies that use assumptions about the image's Probability Density Function (PDF).

The *normalization to the maximum* strategy computes I_p as the average value of the 95th bin of the histogram of the image. In other words, this mean averaging the 5% higher intensity values and use this mean as I_p. Another useful approach is the so-called *integral normalization*, which computes I_p as the sum of all values in the image.

Other approaches involves some a-priori knowledge about the intensity distribution of normal subjects in a certain modality. This is the case of setting I_p to the Binding Potential (BP), a ratio between the intensities at specific and non-specific areas **[Scherfler2005]**.

Finally, more advanced approaches use a general linear transformation of the image:

$$I' = aI + b \tag{3.6}$$

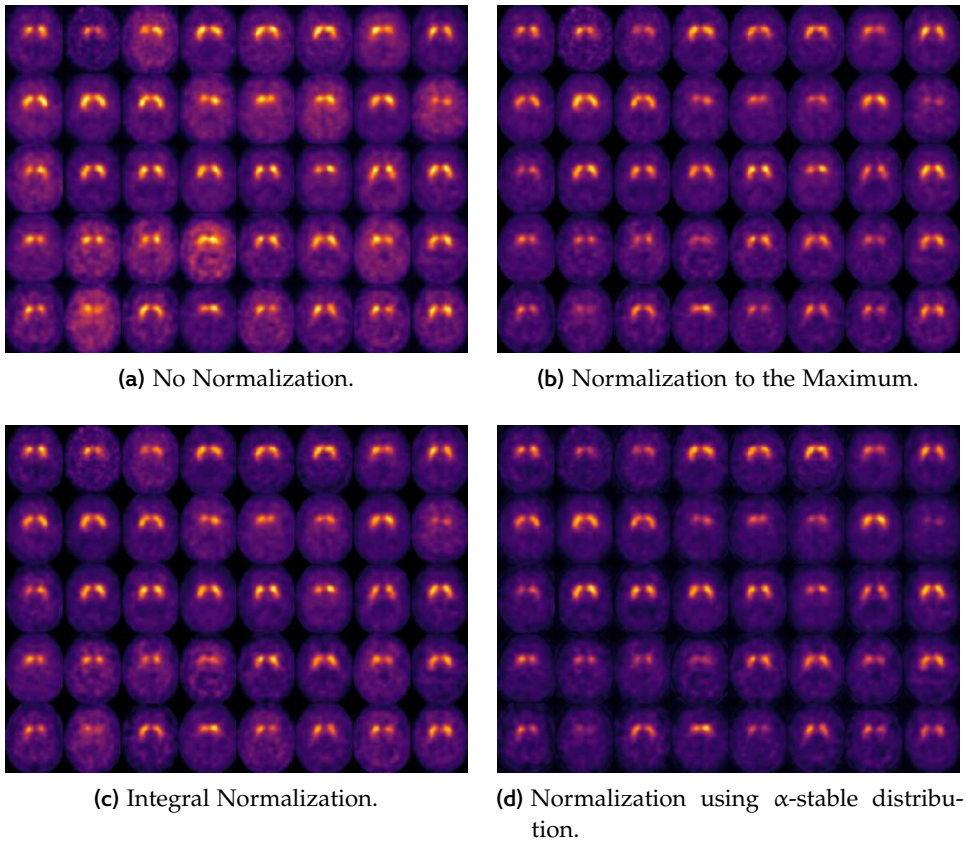
The parameters a and b are so that the PDF of a given matches a reference PDF. There exist methods that use the histogram **[Arndt1996]**, the gaussian distribution or the alpha-stable distribution **[Salas-Gonzalez2013]**. In this latter case, the parameters a y b are computed as linear transformations of some distribution's parameters:

$$a = \frac{\gamma*}{\gamma}, \quad b = \mu* + \frac{\gamma*}{\gamma}\mu \tag{3.7}$$

where $\gamma*$ and γ are the dispersion parameters of the alpha-stable intensity distribution of the non-normalized and the reference image respectively, and $\mu*$ and μ are the location parameters of the same images.

Despite traditionally structural modalities such as MRI did not use intensity normalization, there exist a new tendency towards the use of quantitative T1 - weighted (qT1) and quantitative T2 - weighted (qT2) images **[Weiskopf2013]** that provide biomarkers for absolute measures such as myelination, water and iron levels. This strategy is especially designed to overcome different sources of variability that affect multicentre studies, e.g. magnetic field inhomogeneity, noise, evolution of the scanners, etc. The role of those in multi-centre studies is addressed at Chapter 7.

See Figure 3.3 for a comparison between different strategies of intensity normalization on the same images.



(a) No Normalization.

(b) Normalization to the Maximum.

(c) Integral Normalization.

(d) Normalization using α-stable distribution.

Figure 3.3: Comparison between different types of intensity normalization, applied to the VDLN-DAT dataset (see Appendix A).

3.3 Evaluation Parameters and Methodology

3.3.1 Cross-validation

Some machine learning applications such as digit or faces recognition use tens of thousands of images as input. In these cases, the common practice is to divide randomly the data in three subsets: training, validation and testing [**Bradley1997**]. However, in neuroimaging, sample size is an issue. In contrast to those applications, we only have hundreds of patients in the best case, and the estimation of the performance using these subsets might not be reliable.

In these cases, Cross-validation (CV) is used to obtain more accurate performance measures. CV performs a division of the dataset into several subsets $X = S_1, S_2, ...S_k$ and iteratively use some of these subsets for training or testing.

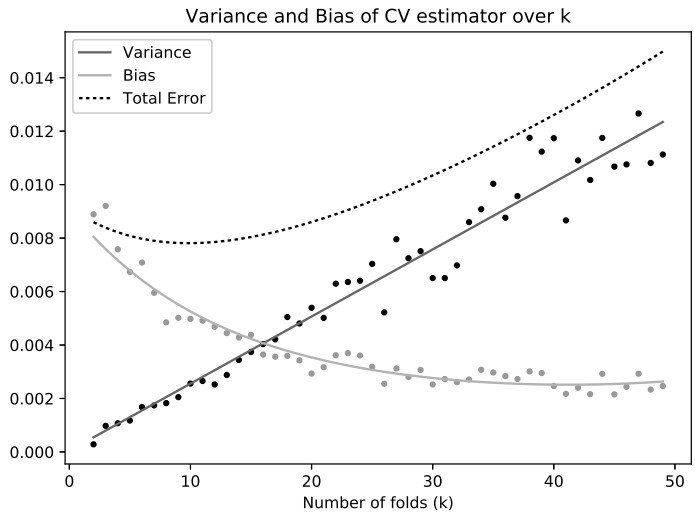


Figure 3.4: Evolution of bias and variance when increasing the number of folds in a k-fold CV.

The simplest CV estimator is k-fold. This approach uses k equally-sized, non-overlapping subset. For each subset S_i (or "fold", hence the name), the model is trained on all subsets $S_k \forall k \neq i$, and then evaluated on S_i. The performance measures, e. g.accuracy, are obtained as the average of the accuracies on each fold.

A particular case where $k = N$ (where N is the exact number of subjects in the dataset) is Leave-One-Out (LOO). This estimator is approximately unbiased for the true accuracy, but can have high variance because there is much overlapping between the N training set [**Hastie2009**]. This imply that the learned models are correlated, and therefore, dependent. All CV strategies with $k > 2$ have overlap, and therefore, high variance. See how variance and bias evolve in a k-fold validation in Figure 3.4.

Using $k = 10$ is assumed as a good compromise between variance and bias in many works [**Kohavi1995**, **Hastie2009**]. In this thesis, when referring to k-fold, we often use stratified cross validation, which is a subclass of k-fold where the distribution of classes within each fold is similar to the distribution of classes in the whole dataset, making the estimates more accurate [**Kohavi1995**].

		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Table 3.1: Confusion matrix and its parts

3.3.2 Classification Performance

From each iteration in the CV loop, a confusion matrix is obtained, from which all performance measures will be obtained.

The confusion matrix (see Table 3.1) accounts for the number of correct and incorrect predictions: True Positives (TPs) and True Negatives (TNs) are correct predictions, and False Positives (FPs) and False Negatives (FNs) are incorrect predictions. It also allow us to identify which type of error is our model making, which in hypothesis testing are known as type I errors(FPs) and type II errors (FNs). The confusion matrix is the basis for computing other performance measures, such as accuracy (acc), sensitivity (sens) or specificity (spec).

$$\text{acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.8}$$

$$\text{sens} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.9}$$

$$\text{spec} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3.10}$$

Sensitivity is also known as TP rate or recall in the literature, and specificity is known as TN rate. Sensitivity is widely used in the medical literature, since it gives an idea of how "sensitive" is our model to the patterns related to a disease (usually considered the positive).

Part II

REDUCING THE FEATURE SPACE

4 IMAGE DECOMPOSITION

In this chapter, we will focus on those CAD systems that use a combination of an image decomposition method and feature selection by means of hypothesis testing. These variety of methods have been published in [**Martinez201141**, **Martinez-Murcia20129676**, **Martinez-Murcia2013255**, **Martinez-Murcia201458**].

Image decomposition methods model a set of samples as a linear combination of c latent variables, also known as components. These variables can be considered as the basis of a c-dimensional space where each sample is represented by a feature vector of length c. The i-th neuroimage in our dataset can be therefore decomposed as:

$$\mathbf{x_i} = \mathbf{s_0 w_0} + \mathbf{s_1 w_1} + \cdots + \mathbf{s_c w_c} + \boldsymbol{\epsilon} = \mathbf{sW} + \boldsymbol{\epsilon} \tag{4.1}$$

Where s_i is the coordinate (or component score) of the current image in the i-th dimension of the new space defined by all the base vectors $\mathbf{w}_i$ (component loadings), and ϵ is the error of the estimation. Figure 4.1 shows an illustration of the process.

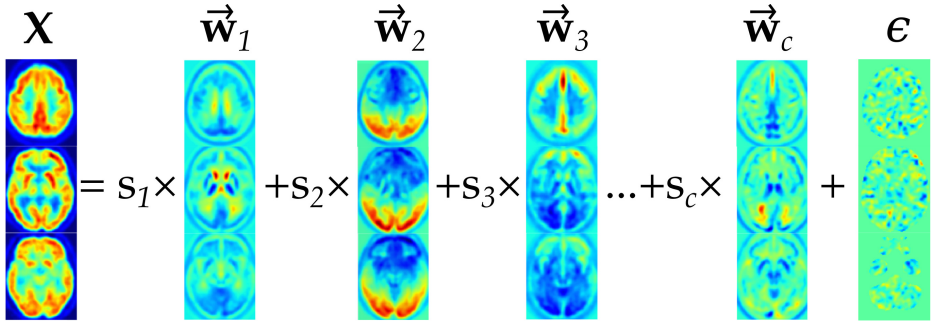


Figure 4.1: Illustration of how decomposition algorithms such as FA and ICA work on a PET-FDG brain image.

Many signal decomposition techniques are used in the literature, for example PCA or PLS [**Spetsieris2009**, **Illan2011**, **Towey2011**, **Segovia2013**, **Khedher2015**]. We will focus on two less known decomposition algorithms Factor Analysis (FA) and Independent Component Analysis (ICA), which we will integrate in different CAD systems using a pipeline similar to the one displayed at Figure 4.2. This

pipeline involves feature selection (for reducing the dimensionality), decomposition of the feature vectors and classification.

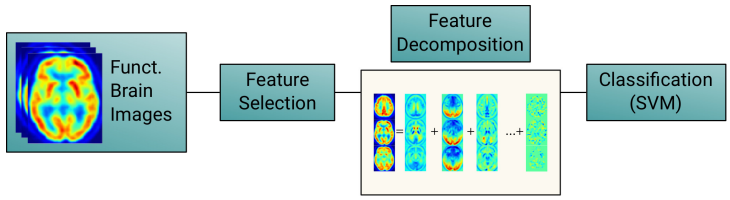


Figure 4.2: Illustration of the system used in Chapter 4.

4.1 Feature Selection

Feature selection is the first strategy used for feature reduction [**Martinez-Murcia2016b**], and it is often used along with feature extraction in order to build more complex pattern recognition systems. It refers to any strategy intended to find a subset of the original features containing the more suitable ones according to a certain criterion. Therefore, irrelevant features are discarded, and resultant models are faster and more cost-effective [**Guyon03**]. However, it usually requires an additional optimization to find the parameters for the optimal feature subset, and furthermore, it is impossible to guarantee that the optimal features for the subset are the same of the full feature set [**DaelemansHosteMeulderEtAl2003**].

In this work, we will use filtering methods to perform feature selection. As we introduced in Section 2.4.2, filtering methods are based on the computation of a feature relevance score directly on the data. The relevance score is used to sort the different features, discarding those with a lower score, and it is usually computed independently for each feature, in what is called a univariate approach [**SaeysInzaLarranaga2007**].

Feature selection can be used before or after feature extraction. When using computationally-intensive algorithms such as FA or especially ICA, the selection of best features prior to the decomposition is key to obtain high performance while keeping the computation times small [**Martinez201141**, **Martinez-Murcia20129676**]. This also removes noise in some cases where the decomposition algorithm cannot compute correctly the variance.

Three feature selection algorithms have been used in this thesis, not only in the CAD systems proposed in this chapter, but in many other models that will be presented later: the t-Test, the Kullback-Leibler divergence or Relative Entropy, and the Mann-Whitney-Wilcoxon rank test.

4.1.1 t-test

The t-test is an old friend of statisticians. In this work we will use the independent two-sample t-test [**Fay10**]. It quantifies the differences between two classes using an assumption of independent variances. Let X_i^f a vector containing the f-th feature of all elements in class i. The t-score of the f-th feature can be computed as:

$$t_f = \frac{\bar{X}_1^f - \bar{X}_2^f}{\sqrt{\frac{\sigma^2_{X_2^f} + \sigma^2_{X_1^f}}{n}}} \tag{4.2}$$

where $\sigma^2_{X_i^f}$ is the variance and $\bar{X}_i^f$ is the average of the f-the feature within class i. The t-test is extensively used in the neuroimaging community, and it is the basis for the SPM and VBM analyses [**spm_book**]. See figure 4.13a for an example of the t-test computed on the ADNI-PET database.

4.1.2 Kullback-Leibler Divergence

Another alternative is the Kullback-Leibler (KL) divergence, also known as Relative Entropy. It is a non-symmetric measure of the difference between two probabilities distributions. Let us assume that X_1^f and X_2^f, the vectors containing the f-th feature of all elements in class i, are two discrete random variables. Therefore, the KL divergence can be calculated with equation 4.3 [**Theodoridis1999**].

$$KL_f = \left(\frac{\sigma^2_{X_2^f}}{\sigma^2_{X_1^f}} + \frac{\sigma^2_{X_1^f}}{\sigma^2_{X_2^f}} - 2 \right) + \frac{1}{2} \left(\bar{X}_2^f - \bar{X}_1^f \right)^2 \left(\frac{1}{\sigma^2_{X_1^f}} + \frac{1}{\sigma^2_{X_2^f}} \right) \tag{4.3}$$

using the same notation than in t-test. See figure 4.13b for an example of the computed KL divergence on the ADNI-PET database.

4.1.3 Mann-Whitney-Wilcoxon

The Mann-Whitney-Wilcoxon (MWW) rank test, also known as U-test, assigns a rank to all values in the vector corresponding to the f-th feature, X^f, without considering any class. The method used to assign a rank is the 'average', which means that each value is assigned with the average of the ranks that would have been assigned to all the tied values. This means that, for example, in the case of the vector $X^f = (0, 2, 3, 2)$, the ranks assigned to each element would be $R^f = (1, 2.5, 4, 2.5)$.

Let n_1 and n_2 be the number of elements in class 1 and 2 respectively, and R^f the vector of ranked elements. We proceed by selecting the first n_1 elements in R^f by:

$$R^f_{n_1} = R^f_i \quad \forall i \in (0, n_1) \tag{4.4}$$

The U-score for the f-th feature and the first class will be:

$$U^f_1 = n_1 n_2 + n_1 \frac{n1+1}{2} - \sum R^f_{n_1} \tag{4.5}$$

And the it can be computed for the second class as the remainder:

$$U^f_2 = n_1 n_2 - U_1 \tag{4.6}$$

The final U^f can be assigned to either U^f_1, U^f_2 or $\min U^f_1, U^f_2$ [**Fay10**], but the usual approach nowadays is to assign $U^f = U^f_2$. Unlike t-test, MWW test does not assume any prior distribution, and therefore is less likely than it to spuriously indicate significance because of the presence of outliers. Under the normal distribution, it performs relatively similar [**Fay10**]. See figure 4.13c for an example of the MWW U-test computed on the ADNI-PET database.

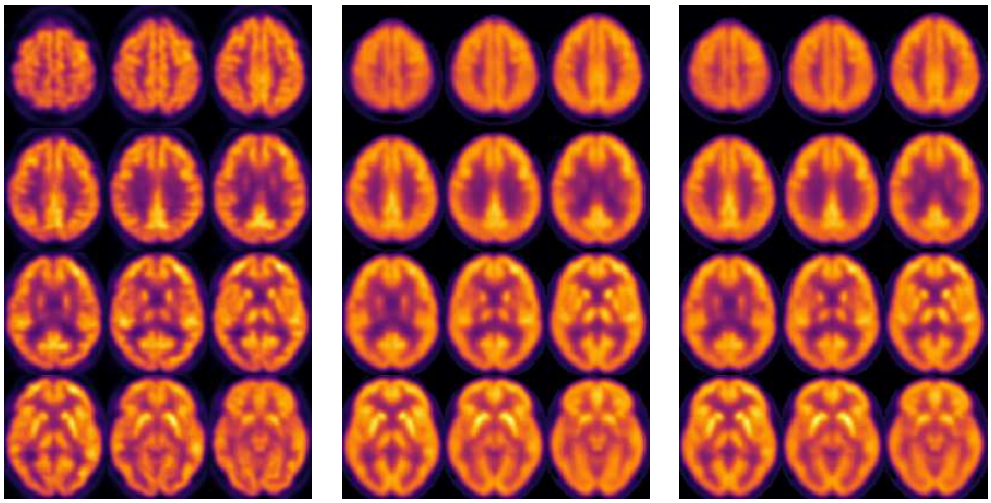
4.2 Decomposition Algorithms

The feature selection algorithms presented above will perform a significant feature reduction, from hundreds of thousands of voxels to a few thousands. These few thousands voxels are considered the best in discriminating between CTL and affected subjects in each of the diseases. The feature selection strategy can be thought of as a mask, in which only the most relevant regions according to the tests are selected (see Figure **??**).

However, this number of features is still large, and therefore, further feature reduction can be applied by performing a decomposition of the masked regions. We have used two algorithms in our CAD systems: Factor Analysis (FA) and Independent Component Analysis (ICA).

4.2.1 Factor Analysis

Factor Analysis (FA) was used in [**Martinez201141**, **Martinez-Murcia20129676**] to perform feature extraction in CAD systems. This strategy assumes that each image in the database is a realization of a given experiment. FA then models each of the N observations (or subjects) as the expression of c unobserved variables, known as factors. The model follows the general decomposition equation



(a) Original **(b)** FA reconstruction **(c)** ICA reconstruction

Figure 4.3: Original PET image from the ADNI-PET dataset, and examples of reconstruction using FA or ICA, with 10 components.

(Eq. 4.1), but assuming that the dataset matrix $\mathbf{X}$ is zero-centred. That is, that we have subtracted the mean prior to the computation. In matrix form, Eq. 4.1 can be rewritten as:

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{SW} + \boldsymbol{\epsilon} \tag{4.7}$$

The columns of $\mathbf{W}$ are known as factors, and the rows of $\mathbf{S}$ are known as loadings (similar to the concept of component loading and component scores in PCA). Thanks to this, we can convert the original dataset $\mathbf{X}$ of size $N \times f$ into $\mathbf{S}$, of size $N \times c$. The procedure of computing the decomposition imposes some assumptions on $\mathbf{W}$:

- $\mathbf{W}$ and $\boldsymbol{\epsilon}$ must be independent.
- $E[\mathbf{W}] = 0$.
- $\text{Cov}(\mathbf{W}) = \mathbf{I}$, which ensures that the factors are uncorrelated.

Now we can rewrite Eq. 4.7 as:

$$\text{Cov}(\mathbf{X} - \boldsymbol{\mu}) = \text{Cov}(\mathbf{SW} + \boldsymbol{\epsilon}) \tag{4.8}$$

Under the previous constraints, and setting $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X} - \boldsymbol{\mu})$, Eq. 4.8 becomes:

$$\boldsymbol{\Sigma} = \mathbf{S}\text{Cov}(\mathbf{W})\mathbf{S}^T - \text{Cov}(\boldsymbol{\epsilon}) \tag{4.9}$$

Since $\mathrm{Cov}(\mathbf{W}) = \mathbf{I}$, and making $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi}$, the diagonal matrix containing the specific variances of the reconstruction error, we obtain the alternative form of FA:

$$\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^{\mathsf{T}} - \boldsymbol{\Psi} \tag{4.10}$$

The mean $\boldsymbol{\mu}$, and the matrices $\mathbf{S}$ and $\boldsymbol{\Psi}$ are obtained via Maximum Likelihood estimation. To guarantee an unique solution, we impose that $\mathbf{S}^{\mathsf{T}}\boldsymbol{\Psi}^{-1}\mathbf{S}$ is a diagonal matrix. Then, we obtain the parameters by maximizing the log-likelihood given by the following expression:

$$\ell(\mu, \mathbf{S}, \boldsymbol{\Psi}) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log\left|\mathbf{S}\mathbf{S}^{\mathsf{T}} + \boldsymbol{\Psi}\right| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)^{\mathsf{T}}(\mathbf{S}\mathbf{S}^{\mathsf{T}} + \boldsymbol{\Psi})(\mathbf{x}_i - \mu) \tag{4.11}$$

FA differs from PCA mainly because it performs an estimation of the noise, and needs the number of factors c as an input. Choosing c is not a naive task. A large c can yield a small reconstruction error, but the factors will not be representative enough, leading to overfitting of the subsequent model. Conversely, a small c can lead to a large reconstruction error, causing information loss. We have computed the reconstruction error variance over the ADNI-PET dataset, and plotted it in Figure 4.4 (similar graphs can be obtained for other databases. This proves that the error is assymptotical as we increase c, and therefore, once arrived at certain error, the improvements are not significant. To observe how the error affects the reconstruction, in Figure 4.3b we can compare a reconstructed image with its corresponding original.

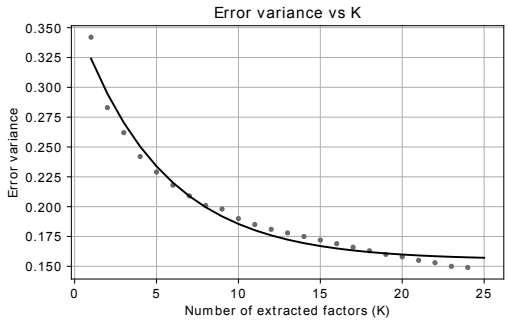


Figure 4.4: Specific variance of reconstruction error Ψ using FA, in function of number of factors extracted (K) for ADNI-PET database (the behaviour is similar in other datasets).

4.2.2 Independent Component Analysis

Independent Component Analysis (ICA) [**Hyvarinen2000**] is an algorithm that performs decomposition imposing that the resulting components must be independent. It was used in [**Alvarez2009**, **Martinez201141**, **Martinez-Murcia20129676**] as part of a CAD system, and it had been used in other medical imaging applications such as segmentation [**DeMartino2007**].

ICA was born as a solution to the *blind source separation* problem, in which the aim is to estimate c independent sources from a series of mixed signals [**Hyvarinen2000**]. To do so, we assume the source signals to be non-gaussian, in addition to the independence assumption that we mentioned before. That is why their authors consider ICA to be a non-gaussian version of FA [**Hyvaerinen2003**], although due to this assumption, the results are very different to those obtained in FA.

Unlike FA, ICA does not account for noise in the estimation procedure, and therefore the equation remains:

$$\mathbf{X} = \mathbf{WS} \tag{4.12}$$

where $\mathbf{S}$ are the component scores and $\mathbf{W}$ are the component loading, 'sources' or 'mixing matrix'. Given that ICA lacks a noise term, there is a procedure called *whitening* that must be applied for the algorithm to converge [**Hyvarinen2000**]. The whitening implies a linear transformation of the i-th observed variable $\mathbf{x}_i$ into a *white* vector $\tilde{\mathbf{x}}_i$ so that its covariance matrix equals the identity:

$$E\{\tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T\} = \mathbf{I} \tag{4.13}$$

This procedure is often performed using the eigen-value decomposition (EVD) of the covariance matrix $E\{\mathbf{x}_i \mathbf{x}_i^T\} = \mathbf{EDE}^T$. $\mathbf{E}$ is the covariance matrix containing the eigenvectors of $E\{\mathbf{x}_i \mathbf{x}_i^T\}$, and $\mathbf{D}$ is a diagonal matrix whose diagonal elements are the eigenvalues of $E\{\mathbf{x}_i \mathbf{x}_i^T\}$. Whitening is done using the following equation:

$$\tilde{\mathbf{x}}_i = \mathbf{ED}^{-1/2}\mathbf{E}^T\mathbf{x}_i \tag{4.14}$$

This procedure transform the mixing matrix to:

$$\tilde{\mathbf{x}}_i = \mathbf{ED}^{-1/2}\mathbf{E}^T\mathbf{W}\mathbf{s}_i \tag{4.15}$$

which is indeed orthogonal, as can be seen here:

$$E\{\tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T\} = \tilde{\mathbf{W}} E\{\tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i^T\} \tilde{\mathbf{W}}^T = \tilde{\mathbf{W}}\tilde{\mathbf{W}}^T = \mathbf{I} \tag{4.16}$$

This property reduces the number of parameters to be estimated, since an orthogonal matrix contains $n(n-1)/2$ degrees of freedom, in contrast to the n^2 degrees of freedom of the original mixing matrix $\mathbf{W}$.

Thanks to the central limit theorem, we assume that the sum of a large number of independent random variables tends will be approximately normally distributed, regardless of the individual statistical distributions [**Rice2006**]. This property is used to maximize non-gaussianity and independence in the sources using any independence criteria such as the kurtosis or negentropy in any of the proposed algorithms. In this work, we will use the FastICA algorithm.

4.2.2.1 *FastICA*

FastICA is a block fixed-point iteration algorithm [**Oja1997**, **FastICA99**] based on negentropy as a non-gaussianity measure. Fixed-point algorithms are converge faster than adaptive algorithms [**FastICA99**]. The FastICA algorithm can be considered a neural algorithm [**Hyvarinen2000**], where the weight vector $\mathbf{w}$ can be updated using a learning rule. FastICA defines a learning rule that finds a direction $\mathbf{w}$, a unit vector such that the projection $\mathbf{w}^T\mathbf{x}_i$ maximizes non-gaussianity [**FastICA99**].

The non-gaussianity measure used here is the negative entropy, or negentropy. The negentropy is a form of differential entropy, which for a random vector $\mathbf{y}$ is defined as:

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y}) \tag{4.17}$$

where $\mathbf{y}_{gauss}$ and $\mathbf{y}$ share the same covariance matrix, although $\mathbf{y}$ is not a gaussian random variable, and $\mathbf{y}_{gauss}$ is. There are many approximations to negentropy. The FastICA defines negentropy using the function:

$$J(y) \propto [E\{G(y)\} - E\{G(v)\}]^2 \tag{4.18}$$

where we assume that y is of zero mean and unit variance, v is a Gaussian variable sharing the same mean and variance, and $G(x)$ is any non-quadratic function. Many functions have been proposed, but in the FastICA algorithm we use either $G(x)_1 = (1/a_1)\log\cosh a_1 x$ with $1 < a_1 < 2$ or $G(x)_2 = \exp(-x^2/2)$ [**FastICA99**].

With these measures, we can compute the derivatives of these functions by:

$$g_1(x) = \tanh(a_1 x), \tag{4.19}$$

$$g_2(x) = x\exp(-x^2/2) \tag{4.20}$$

The algorithm for the one-unit version of FastICA can be defined [**FastICA99**] as:

1. Choose an initial (e.g. random) weight vector $\mathbf{w}$.
2. Let $\mathbf{w}^+ = E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}$
3. Let $\mathbf{w} = \mathbf{w}^+/\|\mathbf{w}^+\|$
4. If not converged, go back to 2.

The algorithm considers that the values of $\mathbf{w}$ converge when their dot product is close to 1, that is, they are pointing in the same direction. Note that the expectations are computed as the sample mean in the FastICA algorithm. Additional modifications were presented in [**Hyvarinen2000**], in which step 2 is converted to a Newton iteration and further simplification is performed.

This is the algorithm for one computational unit, or neuron, which computes one component. However, the procedure can be extended to c components by defining c neurons with weight vectors $\mathbf{w}_1, ..., \mathbf{w}_c$ so that $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_n)^T$. The outputs $\mathbf{w}_1^T\mathbf{x}, ..., \mathbf{w}_n^T\mathbf{x}$ must be decorrelated to prevent them from converging to the same maxima, using three methods proposed in [**Hyvarinen2000**].

The method used in this work uses a two-step iterative algorithm [**Hyvarinen2000**] to decorrelate the outputs after each iteration:

1. Let $\mathbf{W} = \mathbf{W}/\sqrt{\|\mathbf{W}\mathbf{W}^T\|}$.
2. Let $\mathbf{W} = \frac{3}{2}\mathbf{W} - \frac{1}{2}\mathbf{W}\mathbf{W}^T\mathbf{W}$

And repeat step 2 until convergence. For simplicity, the norm in step 1 can be computed as any norm but the Frobenius norm, for example, the L2-norm or the largest absolute row sum.

4.3 Results

In this work we will analyse the behaviour of the system proposed in the introduction and illustrated at Figure 4.2. The system comprises the selection of the most relevant voxels using filtering methods (we will focus on t-test, relative entropy and wilcoxon) and a feature decomposition of these using either FA or ICA. Finally, the feature vectors are classified using a SVC with linear kernel, and performance values are obtained via cross-validation (see Section 3.3 for more information).

We vary the number of selected voxels and the number of factors or components depending on the algorithm and the dataset used and evaluate the system with those characteristics. That way, we obtain an estimation of the performance of the system in different situations, so that we can draw conclusions on the disease patterns and the ability of the system in the detection of different diseases.

4.3.1 Alzheimer's Disease

We begin by applying the proposed feature selection plus decomposition pipeline to the two functional neuroimaging datasets: ADNI-PET and VDLN-HMPAO. For this experiment we will use a maximum of 20,000 selected voxels and 25 components.

4.3.1.1 *Factor Analysis*

First, we use FA as a decomposition technique. In Figure 4.5 we average the accuracy over the number of voxels or the number of components respectively, to look at how these variables affect the performance of the system, and we do this for the three filtering methods used.

We can observe that the results are generally better when using the ADNI-PET dataset than with the VDLN-HMPAO, and this is especially notorious when using the relative entropy selection criterion. The performance tends to slightly increase with the number of voxels selected, but it is not the case with the number of components. By looking at figures 4.5b, 4.5d and 4.5f, it seems that a relatively small number of components (approximately 6) is enough to obtain good performance, and afterwards, the performance holds or even decreases.

4.3.1.2 *Independent Component Analysis*

In this section, we compute the results of applying ICA to the ADNI-PET and VDLN-HMPAO datasets. Figure 4.6 depicts the average accuracy over the number of voxels or the number of components respectively for the different selection criteria.

The case is similar to the one presented in Section 4.3.1.1, where the performance slightly improves when increasing the number of selected voxels. The performance is again better when using the ADNI-PET dataset than with the VDLN-HMPAO, although the behaviour is similar.

The results change when varying the number of components. In this case, although good performance is obtained within the first 5 components in most cases, the model achieves similar performance in both datasets with components between 5 and 10, and then, the estimates diverge. In the VDLN-HMPAO, the performance starts to decrease after this number of components, whereas when using the ADNI-PET dataset, the higher average performance is obtained with $c > 15$, especially in the case of the t-test or the wilcoxon selection criteria).

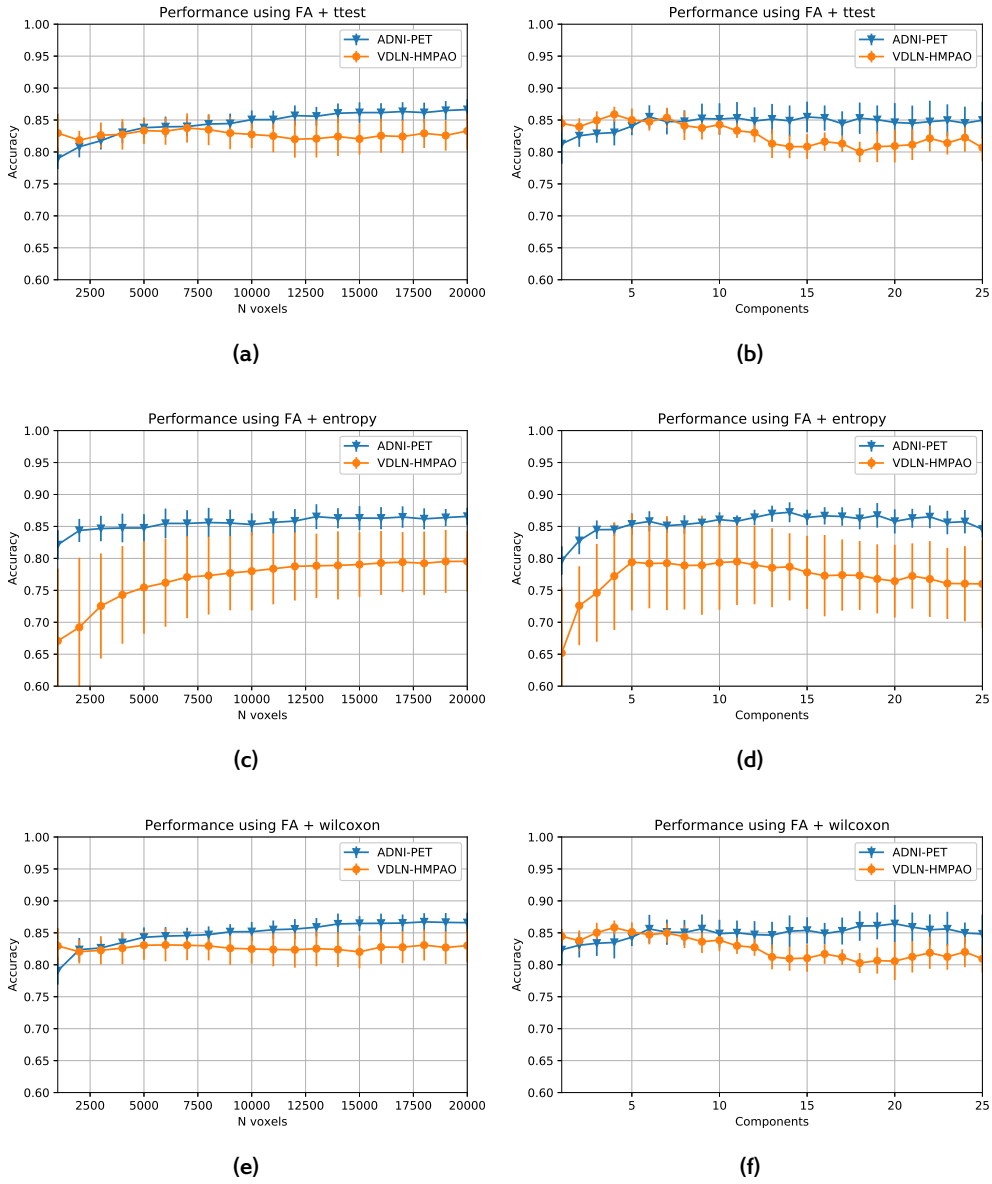


Figure 4.5: Average performance and standard deviation of the proposed system using the two AD datasets, FA and the three feature selection criteria: t-test ((a) and (b)), relative entropy ((c) and (d)) and wilcoxon ((e) and (f)).

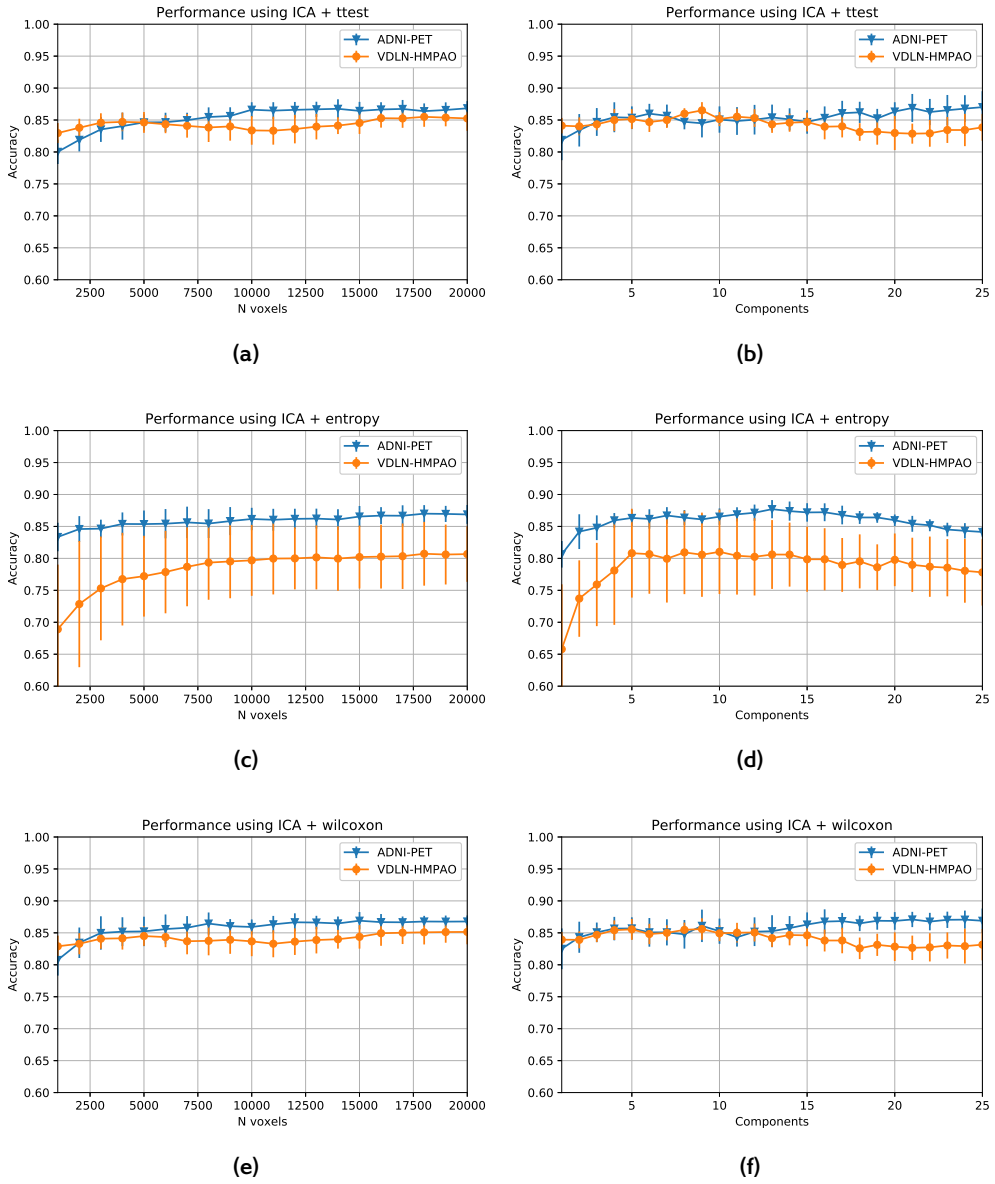


Figure 4.6: Average performance and standard deviation of the proposed system using the three AD datasets, ICA and the three feature selection criteria: t-test ((a) and (b)), relative entropy ((c) and (d)) and wilcoxon ((e) and (f)).

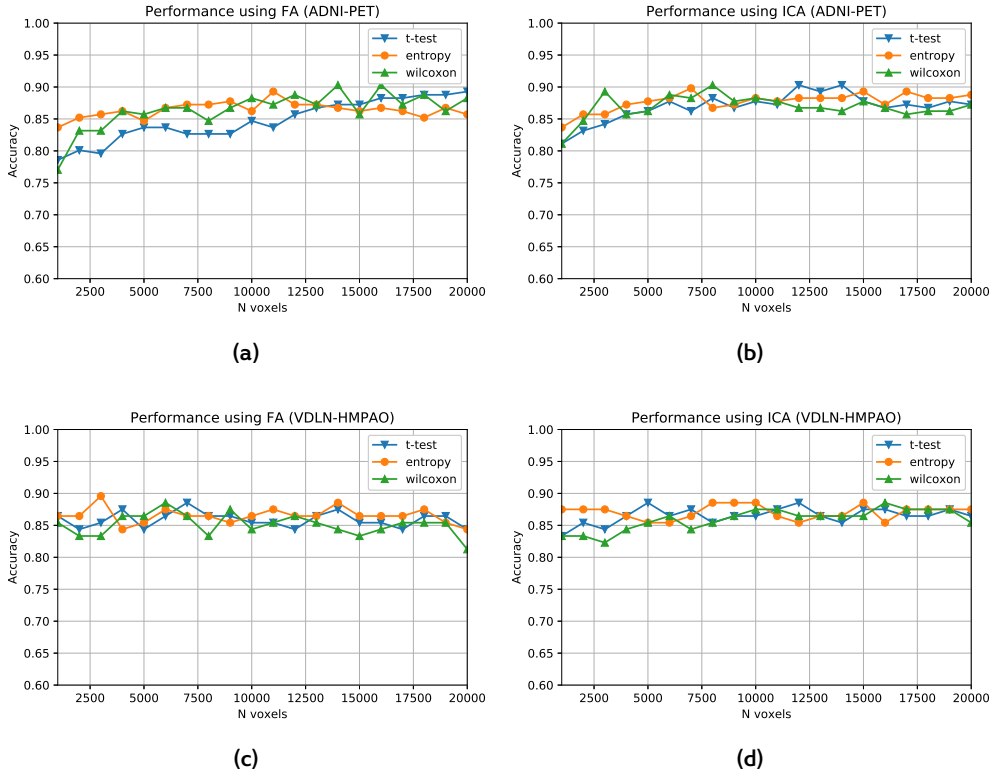


Figure 4.7: Performance of the proposed system using the two AD datasets: ADNI-PET and VDLN-HMPAO at the operation point, and how they vary over the number of selected voxels.

4.3.1.3 *At the Operation Point*

Now we focus on non-averaged values, the values for which our system is optimal: the operation point. In this scenario we see that the tendency is that all systems behave similarly.

When increasing the number of selected voxels, we can see that there is always a tendency of slightly increase in both datasets and decomposition strategies, as can be seen in Figure 4.7. For the ADNI-PET dataset, the maximum accuracies are obtained with a large number of voxels f $>$ 15000, however, in the case of VDLN-HMPAO, we obtain similar performance with fewer voxels, f $<$ 7000 in all selection criteria.

Now we can focus on the performance variations over the number of components in Figure 4.8. The accuracy slightly varies almost in any case, and there is

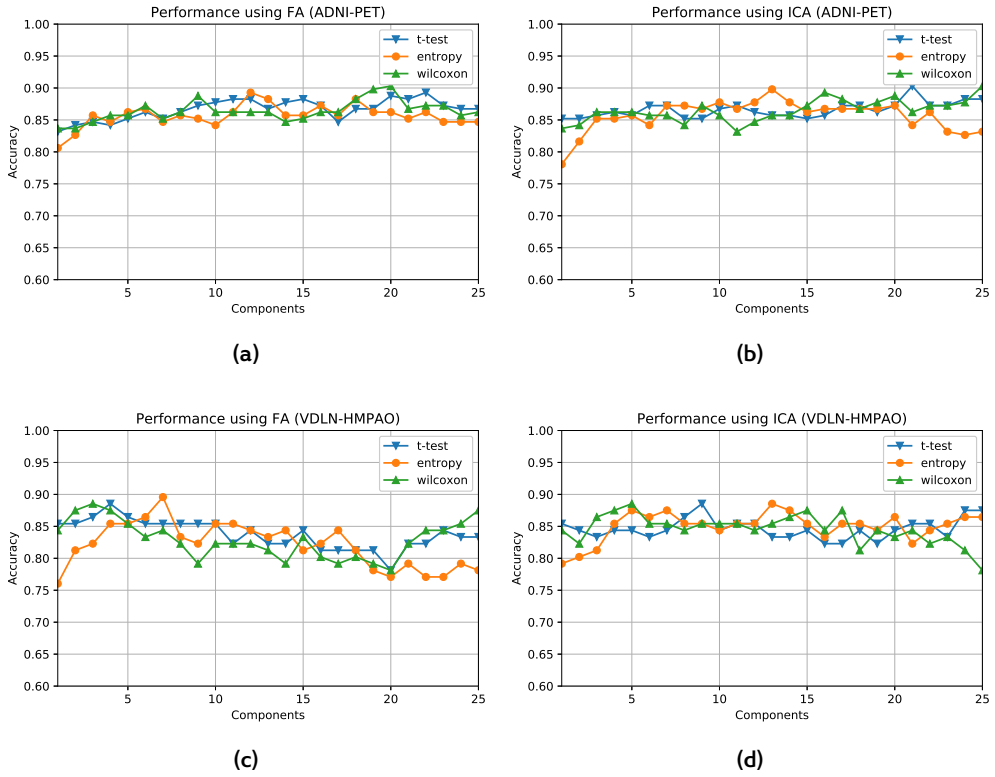


Figure 4.8: Performance of the proposed system using the two AD datasets: ADNI-PET and VDLN-HMPAO at the operation point, and how they vary over the number of components used in the decomposition.

a steep increase in the performance within the first five components in both FA and ICA.

A particular case is the combination of FA and the relative entropy selection criteria applied to the VDLN-HMPAO dataset. In this case there seems to be a trend to achieve a maximum performance at between 5 components. But with the ADNI-PET dataset, the performance keeps and achieves maximum accuracy with $c \approx 20$.

In Table 4.1 we show the performance values obtained at the operation point for our different test combining decomposition algorithms and selection criteria, for the two datasets analysed in this section. It is obvious that both datasets obtain similar performance in almost every case, with values close to 0.9. These values are compared to the baseline classification performance, quantified using VAF [**Stoeckel04**]. From this, we can see that the decomposition systems always

DB	Dec.	Criterion	Accuracy	Sensitivity	Specificity
ADNI-PET	VAF	-	0.882 ± 0.064	0.876 ± 0.099	0.890 ± 0.097
	FA	t-test	0.893 ± 0.074	0.886 ± 0.119	0.901 ± 0.101
		entropy	0.893 ± 0.074	0.894 ± 0.092	0.891 ± 0.088
		wilcoxon	0.903 ± 0.066	0.917 ± 0.079	0.891 ± 0.082
	ICA	t-test	0.903 ± 0.071	0.893 ± 0.100	0.910 ± 0.107
		entropy	0.898 ± 0.059	0.917 ± 0.088	0.881 ± 0.084
		wilcoxon	0.903 ± 0.066	0.906 ± 0.097	0.901 ± 0.094
VDLN-HMPAO	VAF	-	0.802 ± 0.074	0.803 ± 0.088	0.805 ± 0.145
	FA	t-test	0.885 ± 0.076	0.890 ± 0.127	0.875 ± 0.149
		entropy	0.896 ± 0.092	0.907 ± 0.150	0.875 ± 0.139
		wilcoxon	0.885 ± 0.076	0.923 ± 0.130	0.825 ± 0.154
	ICA	t-test	0.885 ± 0.073	0.923 ± 0.130	0.825 ± 0.154
		entropy	0.885 ± 0.076	0.903 ± 0.132	0.850 ± 0.130
		wilcoxon	0.885 ± 0.076	0.907 ± 0.130	0.850 ± 0.152

Table 4.1: Accuracy, sensitivity, specificity, and their standard deviation at the operation point for each method and its corresponding feature selection criterion, using two AD datasets.

perform better than the baseline, but this difference is especially large in the case of the VDLN-HMPAO dataset, which contains very noisy images.

Overall, all methods achieve similar values of accuracy, sensitivity and specificity. When analysing the ADNI-PET dataset, the wilcoxon selection criteria seems to outperform the rest especially with ICA, with a higher accuracy and sensitivity. On the other hand, with the VDLN-HMPAO dataset, either the t-test or the wilcoxon achieves higher sensitivity, but there is no difference in accuracy when decomposing with the ICA algorithm. When using FA decomposition, the relative entropy seems to perform better. In general, there seems to be little difference among methods, and a curious relationship between the relative entropy selection and the VDLN-HMPAO dataset that we will discuss later.

4.3.2 Parkinson's Disease

Now we will look at how the proposed CAD system behaves when applied to the three DaTSCAN datasets: VDLN-DAT, VDLV-DAT and PPMI-DAT. For this experiment we will use a maximum of 1500 selected voxels and 25 components, and images that have been previously intensity normalized using the integral normalization algorithm (see Section 3.2).

4.3.2.1 *Factor Analysis*

Firstly, we will explore the average behaviour of the system that uses FA as a decomposition technique. For this purpose, as in previous sections, Figure 4.9 shows how the average performance varies when varying the number of voxels selected and the number of components extracted.

In this case there is a clear difference between datasets, since some of them are more complex than others, usually due to a typical acquisition procedure in DaTSCAN. In the VDLN-DAT, the images were often composed only of a few cuts around the striatum, whereas in PPMI and VDLV-DAT this rarely happens. That would explain the average outperformance of these two datasets over the VDLN-DAT in almost all cases.

Their behaviours are consistent. Usually, there is no variation with the number of voxel selected (except in the obvious case of the VDLN-DAT dataset and the relative entropy selection criterion). However, there are repeated trends regarding the number of components or factors used in the computation. We can see that the performance increases in the first components, and once we have achieve a decent number (between 4 and 6), the performance starts to decrease. This could mean that the decomposition in more than 5 or 6 components only introduces noise and leads to a wrong decomposition.

Again, this behaviour is consistent with the VDLN-DAT dataset, except for one case. It does seem that the interaction between the VDLN-DAT dataset and the relative entropy selection criterion leads to a wrong model. We will explore this question later, in the discussion.

4.3.2.2 *Independent Component Analysis*

Regarding the application of the ICA decomposition model to the DaTSCAN datasets. In Figure 4.10 we present the average accuracy achieved by this model using three different selection criteria and the three PKS datasets.

In average, PPMI-DAT and VDLV-DAT achieve similar performance, while VDLN-DAT generally achieves much poorer results. However, the behaviour of the system when varying the number of voxels or number of components is similar in all three datasets. The tendency is that accuracy does not significantly vary when increasing the number of voxels (except in the case of the relative entropy and VDLN-DAT). In contrast, when varying the number of components, we observe that the maximum performance is achieved in the first components, typically between 5 and 10.

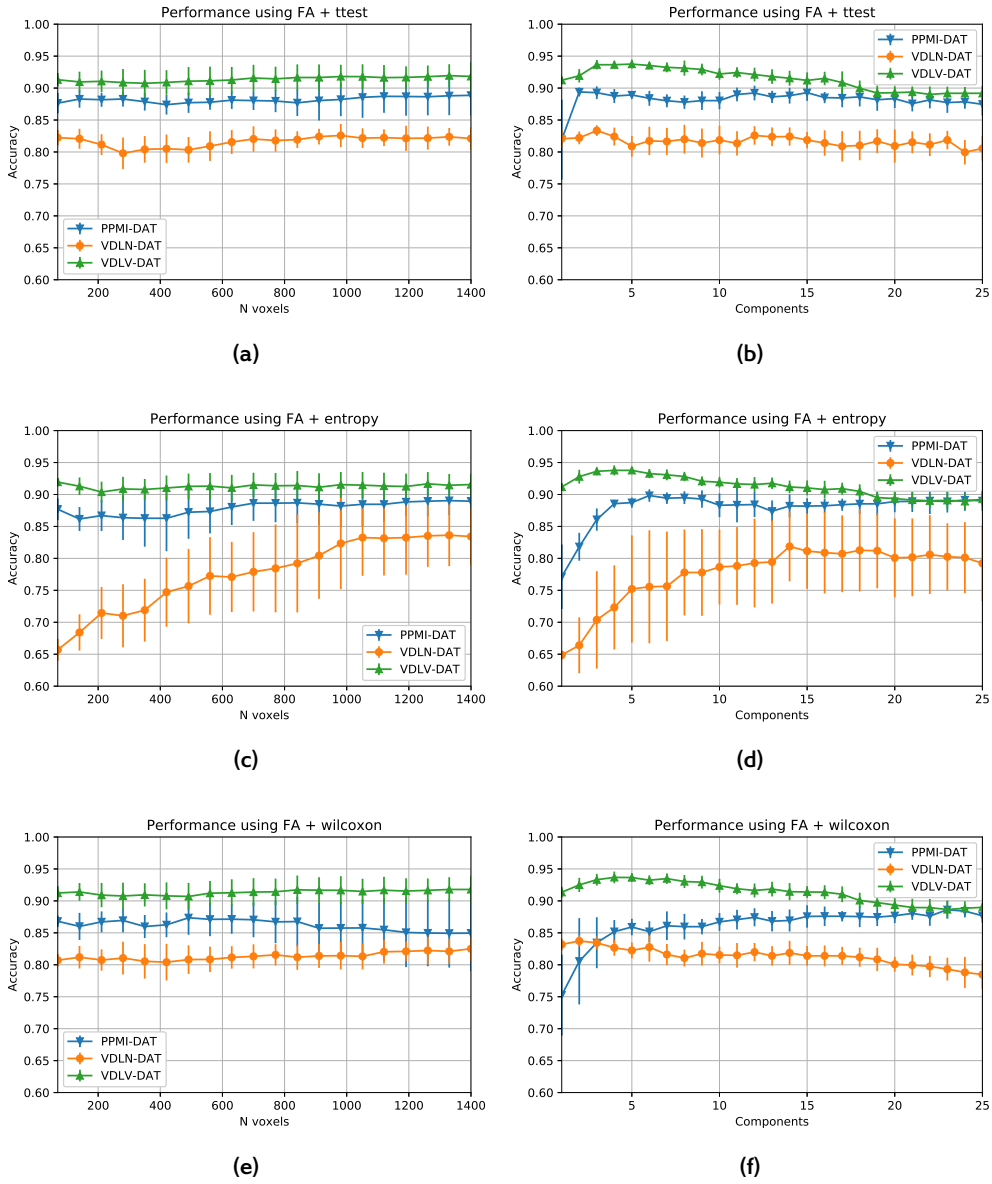


Figure 4.9: Average performance and standard deviation of the proposed system using the three PKS datasets, FA and the three feature selection criteria: t-test ((a) and (b)), relative entropy ((c) and (d)) and wilcoxon ((e) and (f)).

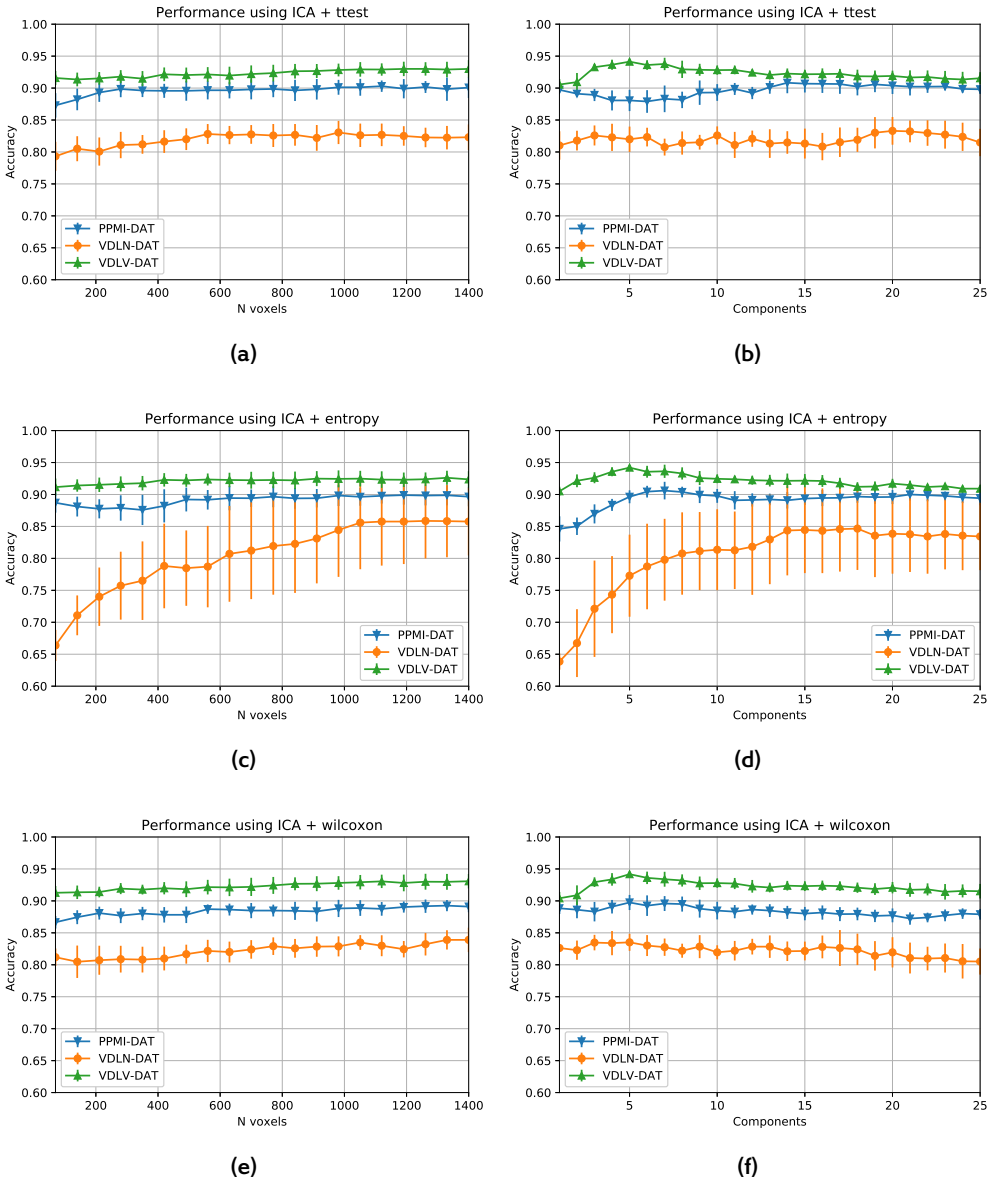


Figure 4.10: Average performance and standard deviation of the proposed system using the three PKS datasets, ICA and the three feature selection criteria: t-test ((a) and (b)), relative entropy ((c) and (d)) and wilcoxon ((e) and (f)).

4.3.2.3 *At the Operation Point*

Let us now have a look at the behaviour of this system at the operation point, using the parameters c and f for which our system is optimal. When varying the number of voxels selected, we obtain the graphs presented at Figure 4.11.

In this case, we obtain again that our system maintains approximately the same performance regardless of the number of voxels selected when tested on both the PPMI-DAT and the VDLV-DAT datasets. In contrast, the performance increases when increasing the number of selected voxels when testing the VDLN-DAT dataset, especially when using the relative entropy criterion. This latter dataset also obtains less performance, whereas the VDLV-DAT achieves the best.

As for the changes in performance when varying the number of components, the results for both FA and ICA based systems are shown in Figure 4.12.

In this first case, the most evident result is the general performance of our system in the three datasets. It is clear that the system performs better when tested on VDLV-DAT than when tested on PPMI-DAT, and both datasets outperform, VDLN-DAT. It is even clearer that when testing on VDLV-DAT, the results are similar using any type of decomposition and selection criteria, with similar performance. We will discuss this issue in more detail later.

The tendency of the performance at the operation point is similar to the average behaviour commented before. In general, there is an accuracy increasing in the first components (usually, between 5 and 10 depending on the decomposition) and then, the performance remains stable. Again, the combination of relative entropy selection and VDLN-DAT achieves striking results. For this dataset, the higher performance is obtained with more than 10 components ($c = 14$), and shows higher variability than other datasets.

Now we will focus on the specific performance values obtained at the operation point, that can be seen in Table 4.2. In this table we observe the differences in performance between datasets and also between CADs using each decomposition strategy.

In general, the systems using ICA tend to perform slightly better than those using FA, although the difference is small. There is little difference between selection criteria as well, although the combination of relative entropy and ICA seems to work better, at least in the PPMI-DAT and VDLN-DAT (in VDLV-DAT all combinations perform equally well).

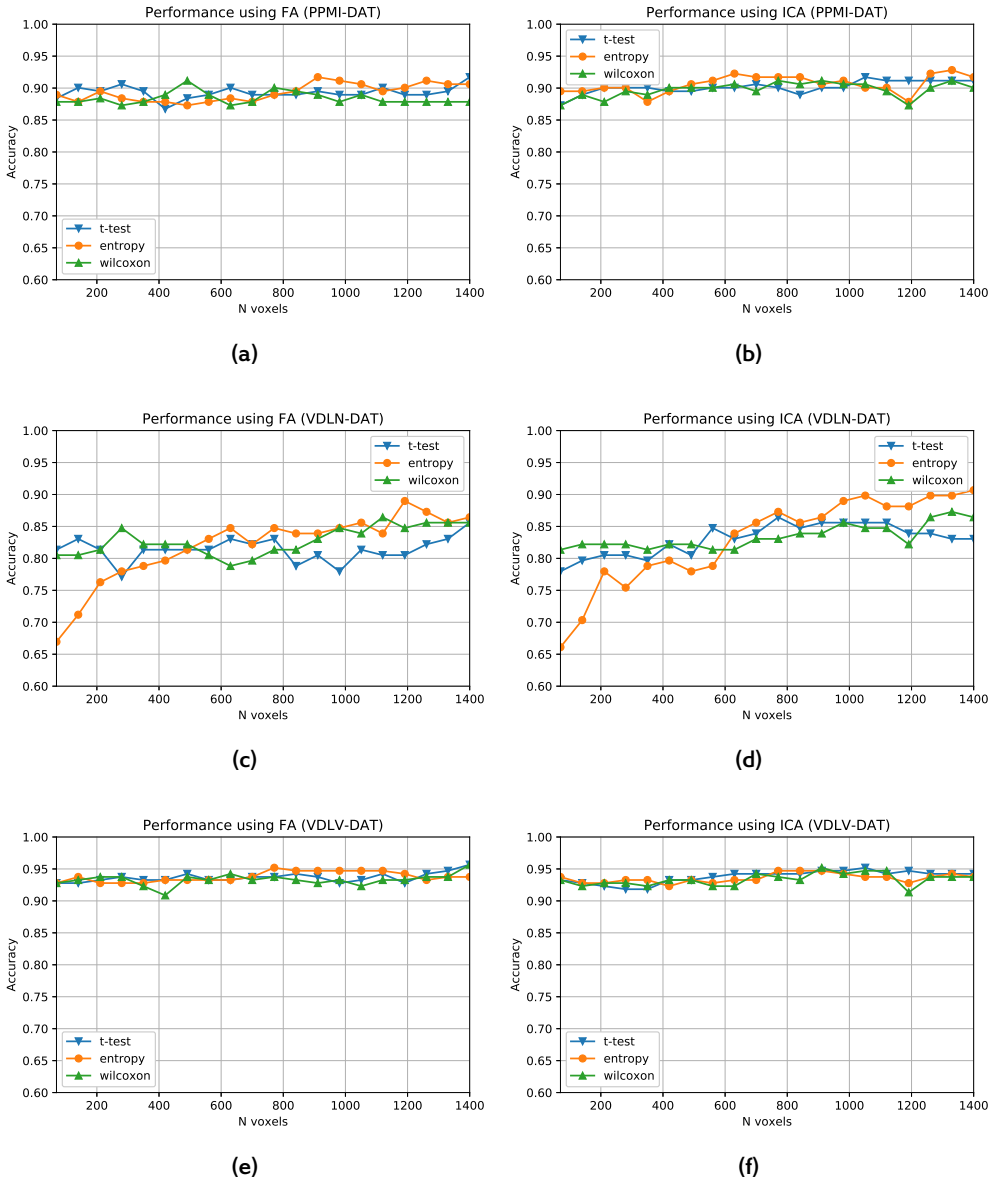


Figure 4.11: Performance of the proposed system using the two PKS datasets: PPMI-DAT, VDLN-DAT and VDLV-DAT at the operation point, and how they vary over the number of selected voxels.

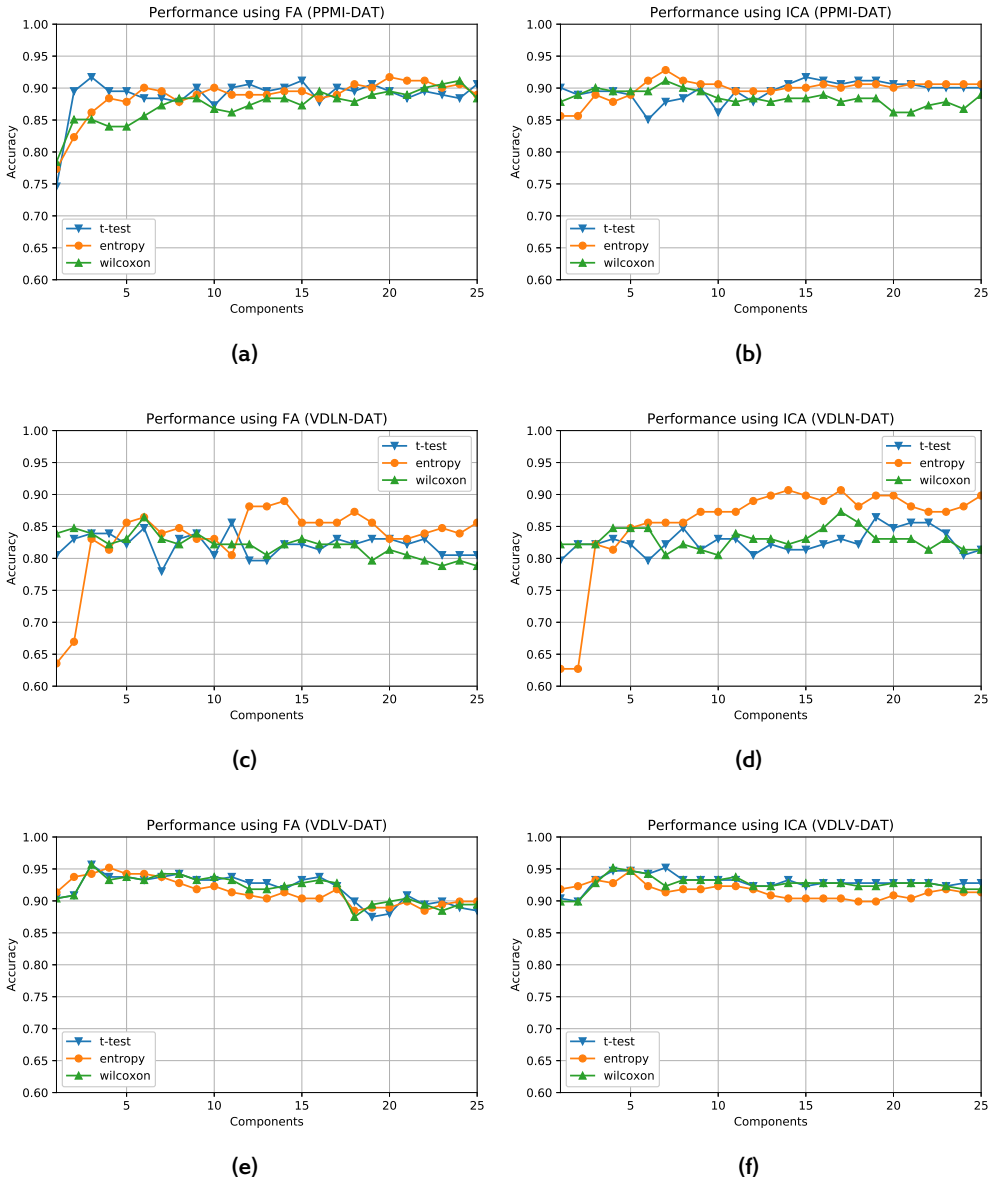


Figure 4.12: Performance of the proposed system using the two PKS datasets: PPMI-DAT, VDLN-DAT and VDLV-DAT at the operation point, and how they vary over the number of components used in the decomposition.

DB	Dec.	Criterion	Accuracy	Sensitivity	Specificity
PPMI-DAT	VAF	-	0.800 ± 0.071	0.831 ± 0.093	0.747 ± 0.112
	FA	t-test	0.917 ± 0.037	0.918 ± 0.095	0.918 ± 0.091
		entropy	0.917 ± 0.060	0.918 ± 0.076	0.921 ± 0.120
		wilcoxon	0.912 ± 0.056	0.927 ± 0.098	0.889 ± 0.102
	ICA	t-test	0.917 ± 0.056	0.900 ± 0.095	0.948 ± 0.109
		entropy	0.928 ± 0.055	0.909 ± 0.091	0.961 ± 0.090
		wilcoxon	0.912 ± 0.070	0.909 ± 0.100	0.920 ± 0.118
VDLN-DAT	VAF	-	0.796 ± 0.129	0.860 ± 0.143	0.675 ± 0.208
	FA	t-test	0.856 ± 0.111	0.887 ± 0.178	0.795 ± 0.164
		entropy	0.890 ± 0.098	0.875 ± 0.118	0.910 ± 0.116
		wilcoxon	0.864 ± 0.070	0.916 ± 0.114	0.780 ± 0.183
	ICA	t-test	0.864 ± 0.101	0.873 ± 0.174	0.840 ± 0.166
		entropy	0.907 ± 0.075	0.889 ± 0.124	0.935 ± 0.131
		wilcoxon	0.873 ± 0.108	0.859 ± 0.181	0.890 ± 0.151
VDLV-DAT	VAF	-	0.918 ± 0.062	0.900 ± 0.094	0.926 ± 0.087
	FA	t-test	0.957 ± 0.063	0.910 ± 0.094	0.973 ± 0.065
		entropy	0.952 ± 0.037	0.940 ± 0.066	0.964 ± 0.064
		wilcoxon	0.957 ± 0.033	0.940 ± 0.066	0.973 ± 0.065
	ICA	t-test	0.952 ± 0.037	0.940 ± 0.066	0.964 ± 0.064
		entropy	0.947 ± 0.045	0.940 ± 0.066	0.955 ± 0.076
		wilcoxon	0.952 ± 0.037	0.940 ± 0.066	0.964 ± 0.064

Table 4.2: Accuracy, sensitivity, specificity, and their standard deviation at the operation point for each method and its corresponding feature selection criterion, using three PKS datasets

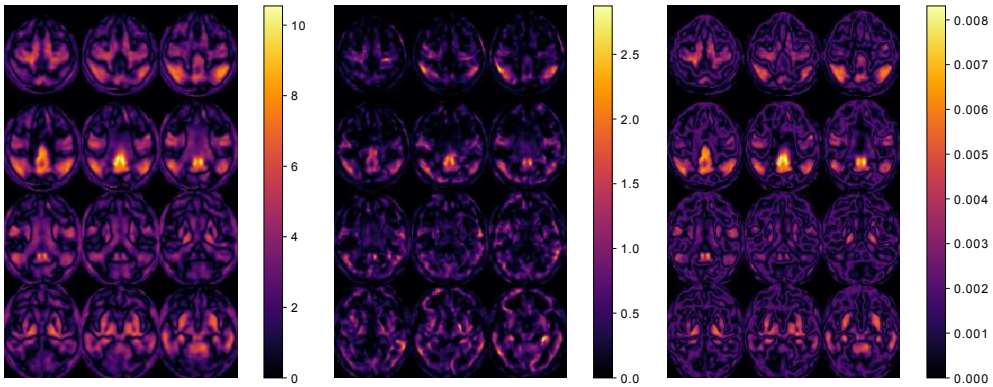
4.4 Discussion

Now we will discuss the general behaviour of the selection and decomposition algorithm in the CAD systems proposed and how they perform on the different diseases and databases.

Our CAD system performs reasonably well on the AD datasets, where it achieves around 90% accuracy, and more than 92% sensitivity. This is achieved in both systems composed by FA and ICA regardless of the selection criterion chosen. The system outperforms the visual analysis estimated by means of VAF [**Stoeckel04**] in both cases. In [**Martinez201141**] and [**Martinez-Murcia20129676**], the systems achieved better performance (up to 95.1% accuracy) when using multivariate quadratic classifiers and ICA, different from the SVC used here.

We have chosen to evaluate the system only on linear SVCs for two main reasons. First, it favours a side-by-side comparison between all methods applied to different datasets in this thesis. And second, linear SVC has been proven to be better generalizable than other systems, even in environments where the small sample size is the norm [**Vapnik1997**].

The selected areas on these CAD systems correspond to the highlighted areas in Figure 4.13, in the case of ADNI-PET dataset.



(a) t-test. **(b)** KL divergence. **(c)** MWW U-test.

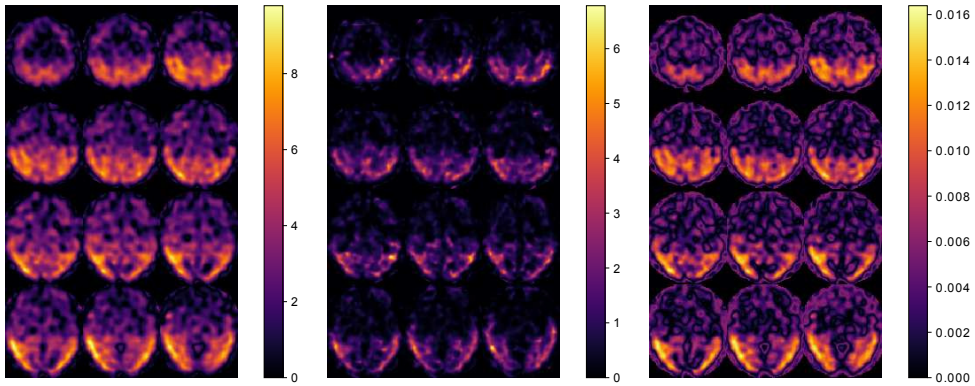
Figure 4.13: Comparison between the different filtering methods, and the regions selected by them, in the ADNI-PET dataset.

In the VDLN-HMPAO different areas are selected as we can see in Figure 4.14. This is mainly due to a change in the modality that deserves to be analysed.

To have a more profound look at the selected regions in both modalities, we provide Table 4.3, where the AAL regions with an overlapping higher than 0.5 in any of the modalities and selection criteria are displayed.

	ADNI-PET			VDLN-HMPAO		
Region	entropy	t-test	wilcoxon	entropy	t-test	wilcoxon
Angular_L	**1.000**	**1.000**	**1.000**	**1.000**	**0.858**	**1.000**
Angular_R	0.729	0.782	**0.809**	**0.980**	0.566	0.697
Cingulum_Post_L	**1.000**	**1.000**	**1.000**	0.589	0.509	0.773
Cingulum_Post_R	**0.843**	**0.852**	0.791	0.524	0.228	0.442
Cuneus_R	0.173	0.121	0.188	**1.000**	0.683	**0.857**
Fusiform_L	0.495	0.156	0.167	0.652	0.561	0.713
Hippocampus_R	**0.807**	0.375	0.413	0.235	0.155	0.425
Occipital_Inf_L	0.378	0.078	0.023	**1.000**	**1.000**	**1.000**
Occipital_Inf_R	0.380	0.080	0.011	**1.000**	0.794	**0.930**
Occipital_Mid_L	0.449	0.184	0.138	**1.000**	**0.881**	**1.000**
Occipital_Mid_R	0.347	0.174	0.128	**0.861**	0.642	0.770
Occipital_Sup_R	0.268	0.094	0.106	**0.955**	0.600	0.736
ParaHippocampal_L	0.778	0.444	0.465	0.098	0.132	0.262
ParaHippocampal_R	**0.917**	0.276	0.316	0.233	0.159	0.311
Parietal_Inf_R	0.497	0.414	0.439	**1.000**	0.339	0.477
Precuneus_L	0.322	0.429	0.463	0.708	0.485	0.660
Precuneus_R	0.311	0.375	0.409	0.770	0.469	0.644
SupraMarginal_L	0.195	0.156	0.183	0.627	0.561	0.762
Temporal_Inf_L	0.748	0.350	0.495	0.676	0.561	0.735
Temporal_Mid_L	0.411	0.244	0.364	**0.952**	0.635	**0.810**
Temporal_Mid_R	0.594	0.234	0.326	**1.000**	0.537	0.711

Table 4.3: Percentage of overlap between the selected areas by each method and the AAL atlas regions. For simplicity, overlapping values higher than 0.8 are displayed in bold.



(a) t-test. **(b)** KL divergence. **(c)** MWW U-test.

Figure 4.14: Comparison between the different filtering methods, and the regions selected by them, in the ADNI-PET dataset.

In the case of the VDLN-HMPAO dataset, the most interesting regions are located at the occipital lobe, the angular lobe and few of them in the temporal lobe. These are selected using almost any of the selection criteria. However, when using the ADNI-PET dataset, the only region with a significant overlapping is the angular lobe, and other regions with a widely documented relation to AD are highlighted [**Dubois2007**, **Claus1994**], such as the cingulum, hippocampus and parahippocampal lobe.

It is clearly noticeable that the relative entropy selection criterion focuses on many different regions, but it is the only one able to detect the hippocampus or parahippocampal lobe in the PET dataset, which other criteria ignore. It also focuses more on the different parts of the occipital lobe in the SPECT dataset. This difference in the selected areas could lead to the different overall performance observed in Figures 4.5c, 4.5d, 4.6c, and 4.6d.

For its part, wilcoxon and t-test often select similar regions. This can be due to their similarity under the normal distribution [**Fay10**], and leads to a higher performance in the systems in average and at the operation point (see Figures 4.5, 4.6, 4.7 and 4.8). From the selected regions, and since t-test and wilcoxon perform generally better, we can infer the more interesting regions for AD classification. For the ADNI-PET dataset, these would be the angular lobes and the cingulum, whereas for the VDLN-HMPAO, we can observe differences in the angular lobe and also all over the occipital lobe and parts of the temporal lobe.

Regarding the decomposition method, there seems not to be any significant differences. Both ICA and FA perform similarly in both datasets, regardless of

the noise contained in the images, although the differences with the baseline system are much higher in the case of the VDLN-HMPAO dataset. This is perhaps due to the smoother nature of the ADNI-PET, in which several images from the same subject were averaged, and therefore, much of the noise was removed, whereas in the VDLN-HMPAO images, the noise could be removed afterwards by discarding many of the lower-significance components.

When applied to the PKS datasets, the results differ from the PPMI-DAT and VDLN-DAT to the VDLV-DAT. These databases differ in the number of subjects that they contain. Whereas in the former there are different subjects with PKS, including subjects without evidence of dopaminergic deficit (SWEDD), in the later we only have PD and CTL subjects, which makes the classification easier.

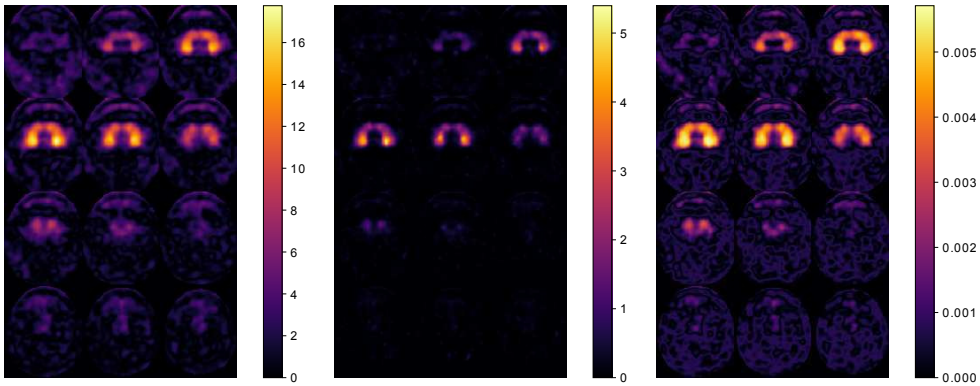
This accuracy differences can be found throughout all figures and tables, although in general, the VDLN-DAT dataset has the lowest performance, while the PPMI-DAT and VDLV-DAT behave similarly in average.

When using FA with the PKS datasets, we observe a similar behaviour to that already seen with AD datasets: there is little difference in performance when varying the number of selected voxels, except when using the relative entropy criterion. In this particular case, there is a rise in the average performance when increasing the number of selected voxels, which is more noticeable when using the VDLN-DAt dataset.

For its part, there are more significant variations when increasing the number of components. Smaller c values lead to a fast increase in performance up to a maximum. Depending on the selection criterion used, this value varies from $c = 3$ when using the t-test to $c = 14$ for the entropy criterion applied to VDLN-DAT. The optimal c is usually located at $c \in [3, 5]$ in most cases, as can be seen in Figure 4.9. A very similar behaviour is achieved when using the ICA decomposition, as Figure 4.10 shows.

In AD we provided a table with the selected regions in both PET and SPECT modalities. Conversely, in DaTSCAN imaging, the selected regions are always in the striatum, and the smaller resolution of these images hardly reveals the underlying structures. The selected regions with either t-test or wilcoxon fundamentally cover the whole caudate, putamen and globus pallidus, and some external structures as well. However, the relative entropy criterion focus almost exclusively in the striatum, with a strong preference for the posterior part, and discards all other regions, introducing less noise.

This is clearly seen in Figure 4.11 where the performance around the operation point is displayed. Here, when looking at the VDLN-DAT dataset, the correlation between performance and number of selected voxels is more obvious. As can be seen in Table 4.2, in the two datasets where SWEDD subjects are included the system which uses relative entropy achieves better results. How-



(a) t-test. **(b)** KL divergence. **(c)** MWW U-test.

Figure 4.15: Comparison between the different filtering methods, and the regions selected by them, in the PPMI-DAT dataset.

ever, in the VDLV-DAT, where the system only involves PD and CTL subjects, the performance is very similar using all three selection criteria.

When looking at the evolution of the performance with the number of selected components, (Figure 4.12), the pattern observed in AD holds for the PPMI-DAT and the VDLV-DAT datasets. In these cases, maximum performance is obtained with a relatively small c (between 4 and 8, depending on the decomposition algorithm). However, with the relative entropy criterion, the VDLN-DAT still needs a higher number (more than 10) to reach the operation point.

All these differences in behaviour could be due to a higher variability in VDLN-DAT, compared to the other two datasets. The number of components needed, especially in FA, points to a more complex decomposition of those images. The sources of variability in this dataset probably correspond to a higher proportion of SWEDD subjects and the number of cuts used in the acquisition. The number of SWEDD in VDLN-DAT is 30 for a total dataset of 148 patients, whereas in the PPMI-DAT dataset we only have 32 SWEDD for 301 subjects. Furthermore, the number of cuts in the images of VDLN-DAT differs from one image to another, since they follow a common practice in which only the ROIs of the brain (the striatum) are acquired.

5 TEXTURE FEATURES

5.1 Introduction

Texture is a household word outside image processing or related fields. However, in that context, it lacks a definition that allow us to measure and quantify it. Patter recognition provides us with a mathematical definition that allow us to use texture as a feature in our CAD systems.

Texture analysis is defined as any procedure by which we can quantify and classify the spatial variation of intensity throughout an image. In neuroimaging, texture has been widely used in segmentation (tissue classification) of MRI images [**Saeed2002**, **Alejo2003**, **Wang2009**], although there exist a number of works using it for feature extraction in CAD-like systems, like the works in [**kovalev2001three**, **sikio2015mr**], or our work on PKS feature extraction [**Martinez-Murcia20**
martinez2014parametrization].

Texture features can be classified in first, second and higher order analysis, depending on the number of variables used. First order statistics [**Martinez-Murcia2016b**] are the most basic form of texture analysis, computing values such as average, variance or histogram of voxel intensity values [**Srinivasan2008**].

The most popular form, with a very developed theoretical background, is second-order statistical texture analysis. This particular form is based on the probability of finding a pair of similar intensities at a certain distance and orientation of a certain image. From these probabilities, many measures can be derived, being the most popular the Haralick texture analysis [**Haralick73**].

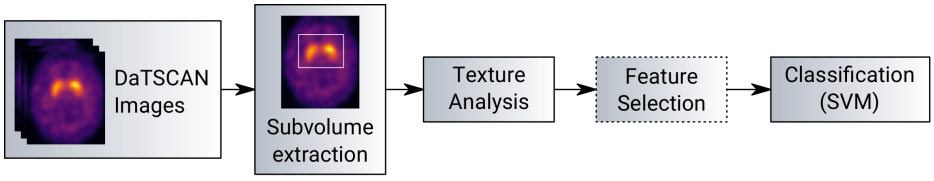


Figure 5.1: Schema of the proposed Texture-based CAD system, including an optional feature selection block.

In this work we have used Haralick texture analysis to extract features from DaTSCAN images and perform an automatic diagnosis of PD. It follows the pipeline depicted at Figure 5.1, as in [**Martinez-Murcia2013266**, **martinez2014parametrizatio**

First we will provide an introduction to the methodology followed at Section 5.2, including the volume selection tools, the Haralick texture analysis and the experiments used to validate the system. Later, in Section 5.3 we define the experiments and show their results. Finally, at Section 5.4 we discuss the implications of this systems and the evaluation results of our texture-based CAD system.

5.2 Methodology

5.2.1 Volume selection

Even the registered DaTSCAN images contain many voxels that are outside the brain. Therefore, to obtain a more robust estimation of the texture, it would be desirable to perform the computation of the features on subvolumes of those images (or subimages) that contain only voxels inside the brain. Many strategies can be performed for this, for example, force the computation of the Grey Level Co-occurrence Matrix (GLCM) to ignore background voxels.

In this work, we opted for extracting a subvolume which contains only voxels higher than a certain intensity threshold I_{th}, which should be specified. To do so, we obtain the maximum and minimum coordinates for which $\mathbf{I}$ is higher than the threshold:

$$p_{x,min} = \arg\min_{x} (\mathbf{I} > I_{th}) \tag{5.1}$$

$$p_{x,max} = \arg\max_{x} (\mathbf{I} > I_{th}) \tag{5.2}$$

And we do the same for the y and z axis of the array. Once this is computed, we can select the volume by:

$$\mathbf{I}_{sub} = \mathbf{I}[p_{x,min} : p_{x,max}, p_{y,min} : p_{y,max}, p_{z,min} : p_{z,max}] \tag{5.3}$$

The resulting subvolume $\mathbf{I}_{sub}$ is the minimum box-shaped volume containing all the values for which $\mathbf{I} > I_{th}$, which allow us to select a I_{th} so that only the regions of interest are contained within.

Different subimages and sizes are obtained when applying different I_{th} In Figure 5.2 we depict a comparison between the resulting images for $I_{th} = [0.25, 0.30, 0.35]$.

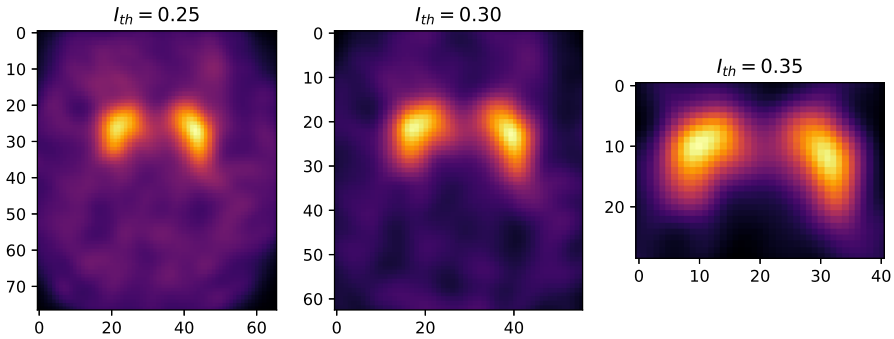


Figure 5.2: Comparison of the different I_{th} values for a random subject extracted from the PPMI database.

5.2.2 Haralick Texture Analysis

5.2.2.1 *Gray Level Co-occurrence Matrix*

The Haralick texture analysis is based on the computation of a Grey Level Co-occurrence Matrix (GLCM), which is a form of evaluating second-order texture statistics. This matrix is a summary of the probabilities of finding a pair of similar grey levels at a certain distance and in a certain direction.

The combination of the unitary vector dimension and the distance defines the offset $\Delta = (d_x, d_y, d_z)$, whose norm is the distance d and is defined in a given spatial direction. In this work, we use a three-dimensional approach to the computation of the GLCM, based on [**Philips2008**], that uses thirteen spatial directions to generalize the standard 2D GLCM to 3D. These offset define different angles and are used to get some degree of rotational invariance [**Philips2008**].

Medical images have different number precision, which can vary from regular 8bit integers (256 values) to the type float64 (1.844×10^{19} possible values). Using all these values, even in the smallest case, would lead to 256×256 matrices, which would be both non representative of the real texture and computationally expensive. Therefore, prior to the GLCM computation, we posterize the image, that is, the image is quantified to use only 16 grey levels. This leads to more tractable GLCM without losing their representativeness.

Once images have been posterized, for two different grey levels i and j, the value of the co-occurrence matrix $\mathbf{C}$ over a $n \times m \times k$ three-dimensional image $\mathbf{I}$ is defined as:

$$\mathbf{C}_{\Delta}(i,j) = \sum_{\mathbf{p}=(1,1,1)}^{(n,m,k)} \begin{cases} 1, & \text{if } \mathbf{I}(\mathbf{p}) = i \text{ and } \mathbf{I}(\mathbf{p}+\Delta) = j \\ 0, & \text{otherwise} \end{cases} \tag{5.4}$$

where Δ is the three dimensional offset that we defined previously, and $\mathbf{p}$ is the position of a given voxel inside the image.

We will compute one 16×16 GLCM for each of the combinations of direction and distances. This matrix $\mathbf{C}_\Delta$ is later modified to create the probability matrix $\mathbf{P}$ as:

$$\mathbf{P}(i,j) = \frac{\mathbf{C}_\Delta(i,j)}{\sum_{i,j} \mathbf{C}_\Delta(i,j)} \tag{5.5}$$

from which the texture features will be derived.

5.2.2.2 *Haralick Texture Features*

In [**Haralick73**, **Haralick1992a**], many texture features are derived from the probability matrix defined above. We have selected twelve of these features to use in this work. These features are:

$$\text{Energy} = \sum_i \sum_j \mathbf{P}(i,j)^2 \tag{5.6}$$

$$\text{Entropy} = \sum_i \sum_j \mathbf{P}(i,j) \log \mathbf{P}(i,j) \tag{5.7}$$

$$\text{Correlation} = \frac{\sum_i \sum_j ij\mathbf{P}(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{5.8}$$

$$\text{Contrast} = \sum_{n=0}^{N_g - 1} n^2 \left\{ \sum_{|i-j|=n} \mathbf{P}(i,j) \right\} \tag{5.9}$$

$$\text{Variance} \sum_i \sum_j (i - \mu_i)^2 \mathbf{P}(i,j) + (j - \mu_j)^2 \mathbf{P}(i,j) \tag{5.10}$$

$$\text{Sum Mean} = \frac{1}{2} \sum_i \sum_j (i\mathbf{P}(i,j) + j\mathbf{P}(i,j)) \tag{5.11}$$

$$\text{Inertia} \sum_i \sum_j (i-j)^2 \mathbf{P}(i,j) \tag{5.12}$$

$$\text{Cluster Shade} \sum_i \sum_j (i + j - \mu_x - \mu_y)^3 \mathbf{P}(i,j) \tag{5.13}$$

$$\text{Cluster Tendency} \sum_i \sum_j \{i + j - \mu_x - \mu_y\}^4 \mathbf{P}(i,j) \tag{5.14}$$

$$\text{Homogeneity} = \sum_i \sum_j \frac{\mathbf{P}(i,j)}{1 + |i-j|} \tag{5.15}$$

$$\tag{5.16}$$

$$\text{Max Probability} = \max_{i,j} \mathbf{P}(i,j) \tag{5.17}$$

$$\text{Inverse Variance} = \sum_{i} \sum_{j} \frac{\mathbf{P}(i,j)}{(i-j)^2} \tag{5.18}$$

where μ_i, μ_j, σ_i and σ_j are the column and row-wise mean and variance respectively. These feature measure things such as the randomness of the grey-level distribution (entropy), the number of repeated pairs (energy), the local contrast or homogeneity of the image, variance, the tendency to form clusters (cluster shade and tendency), among others.

For this work we have used a distance d ranging from 1 to 10, at each of the 13 spatial directions. Therefore, we have computed $13 \times 10 = 130$ GLCMs per image, from which 12 texture features are computed. Our final feature vector will therefore have 1560 features in total.

To further reduce the dimensionality of the feature vector, we have performed feature selection using the t-test, MWW U-test and the relative entropy (KL divergence) criteria (see Section 4.1).

5.2.3 Experiments

For evaluating the system proposed in this chapter, combining texture analysis and other feature selection algorithms, we propose two experiments:

- Experiment 1: Ability of the different texture features to differenciate between PD affected subjects and CTLs. Each of the texture features is analysed in two different ways: a "single approach", which only considers one type of feature using only the matrices at a distance d from the central voxel -and using all the spatial directions- and a "cumulative approach" which considers one type of feature too, but this time using all matrices in distances ranging from 1 to d.

- Experiment 2: Impact of the introduction of a feature selection algorithm after computing the texture features. This allow us to pool all texture features at all distances and directions, and then select the most discriminative ones according to some of these criteria.

All images used are intensity normalized using either normalization to the maximum or integral normalization (see Section 3.2), and afterwards, a subvolume can be extracted using the intensity threshold methodology described at Section 5.2.1. In addition to the feature extraction technique using texture analysis, and the feature selection procedure defined for Experiment 2, we use a

linear SVC for classifying, and 10-fold cross validation strategy (see Section 3.3 for more details).

5.3 Results

5.3.1 Experiment 1

In this experiment, the influence and effect of each texture feature is tested, as in [**Martinez-Murcia2013266**]. We have tested the computation of the GLCMs over the image subvolumes using different thresholds I_{th} (see Sec. 5.2.1) ranging from 0 to 50% of the maximum intensity value, and a range of distances $d = 1, 2, \ldots 10$ in the thirteen spatial directions.

To check which value of the intensity threshold is the best for computing the texture features, we can compute the general tendency of the system by averaging the accuracy values. Figure 5.3 depicts the general trend of the performance over the intensity threshold for either the no normalized or normalized images. This is done for the single and cumulative approach.

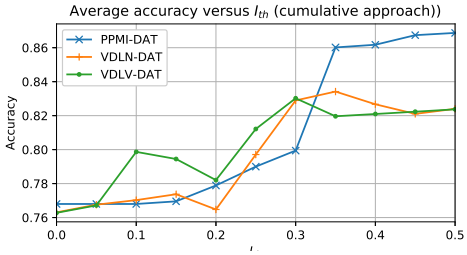
The most obvious differences can be found between normalization procedures. As a general trend, the integral normalization barely has an impact over the performance achieved with the registered images. Furthermore, the normalization to the maximum even drops the performance for high I_{th}, however it increases the general performance in the range 0.1-0.3. From these graphs it is patent that normalization has no impact on the performance achieved by our system, which can be consider an advantage, since it reduces the preprocessing needed.

In this regard, the VDLV-DAT dataset has an strange behaviour. It performance holds and even increases when using normalization to the maximum, but significantly drops when integral normalization is used. This is exactly the opposite as happens to the other dataset, and will be discussed later.

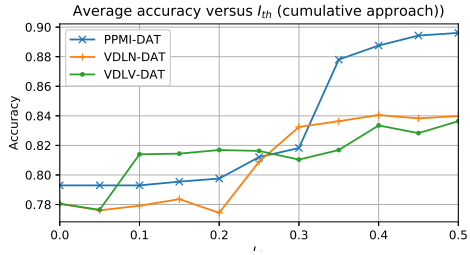
In these images, we can observe a strong dependence of the system's performance with the intensity threshold I_{th}. In general, the performance increases with a more restrictive threshold (a smaller box around the striatum). This increase is probably due to removing the background from the texture analysis. In most cases, $I_{th} \approx 0.35 \times I_{max}$ seems to be a critical value: either the global maximum or the inflection point from which accuracy stabilizes.

the general trend

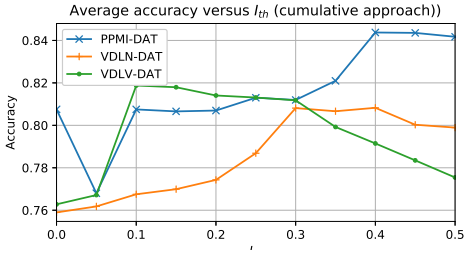
The effect of removing the background is clearly shown in these pictures, obtaining best results when using a $I_{th} > 0.30 \times I_{max}$ and then increasing the accuracy. Furthermore, the effect of the normalization is also clear in these



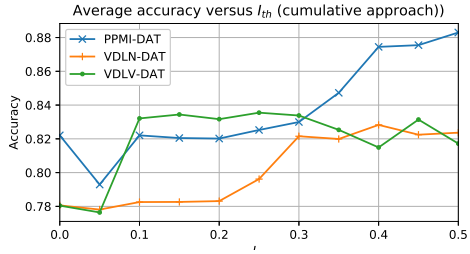
(a) No Normalization (Single)



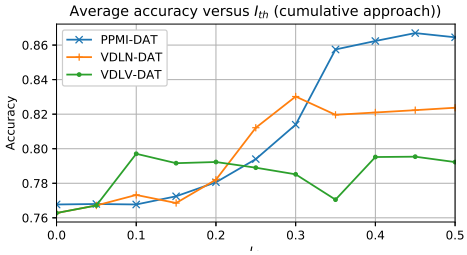
(b) No Normalization (Cumulative)



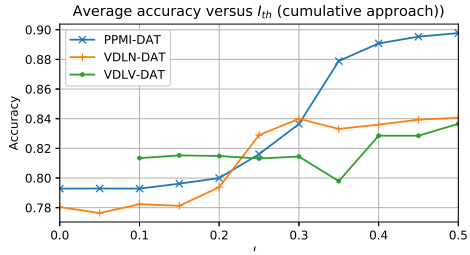
(c) Maximum (Single)



(d) Maximum (Cumulative)



(e) Integral (Single)



(f) Integral (Cumulative)

Figure 5.3: Evolution of the average accuracy values obtained for the single approach and the cumulative approach over the intensity threshold, using no normalization, normalization to the maximum and integral normalization.

two images. It is possible to notice that, when using no-normalized databases (those noted with a "-no" sufix) there are wide ranges of I_{th} values in which similar performance values are obtained, while when using the normalized images, there are obvious peaks of accuracy around some values (usually, 0.30 or 0.35). Having applied no normalization to the images, the average image I_{mean}, from which the subvolume coordinates are extracted, is highly affected by the difference of intensities of different anatomical areas. Therefore, the subvolume computed will not be optimum, and for every value of I_{th} there will be an set of samples in which the texture features will be correctly computed, and another set in which those will be poorly extracted.

The behaviour of each of the Haralick texture features can also be analysed using a box plot (see Section 3.3), to show both numerical accuracy values and the properties (robustness, parameter independence) of using each one. Figure 5.5 depicts all 130 accuracy results of the "single approach" for each feature extracted from a subimage that uses $I_{th} = 0.30 \times I_{max}$ at each distance d (ranging from 1 to 10) in each of the 13 spatial directions. In this case, we can observe that best performance is obtained with the Cluster Tendency in all databases. Good values are also achieved using Homogeneity, Contrast and Correlation. This behaviour is consistent along all three databases, which allow us to propose Cluster Tendency as the best feature to characterize PD patterns.

Finally, to characterize the ability of this single-feature approach, we show its performance at the defined operation point (Using an $I_{th} > 0.30 \times I_{max}$ and a value of $4 < d < 8$) in Table 5.1.

Database - approach	I_{th}	d	Feature	Acc	Sens	Spec	PL	NL
PPMI - cumulates	30	8	Cluster Tendency	0.952	0.973	0.937	15.37	0.029
PPMI - distances	30	6	Cluster Tendency	0.952	0.964	0.943	16.92	0.038
VDLN - cumulates	30	6	Cluster Tendency	0.906	0.911	0.904	9.50	0.098
VDLN - distances	30	6	Cluster Tendency	0.907	0.911	0.904	9.50	0.098
VDLV - cumulates	35	7	Cluster Tendency	0.899	0.879	0.920	10.99	0.130
VDLV - distances	35	6	Cluster Tendency	0.923	0.907	0.940	15.12	0.098

Table 5.1: Accuracy values obtained at the operation point, using Cluster Tendency as a feature. The I_{th} used to compute the GLC matrix is also displayed.

Obviously, the best approach is the cumulative one, given that it contains a bigger amount of information, and thus, describing in a more accurate way the different images. Note that for the cumulative approach, all values of Cluster Tendency computed between $d = 1$ and d are used, while for the single approach, only values of Cluster Tendency at d are considered. However, the

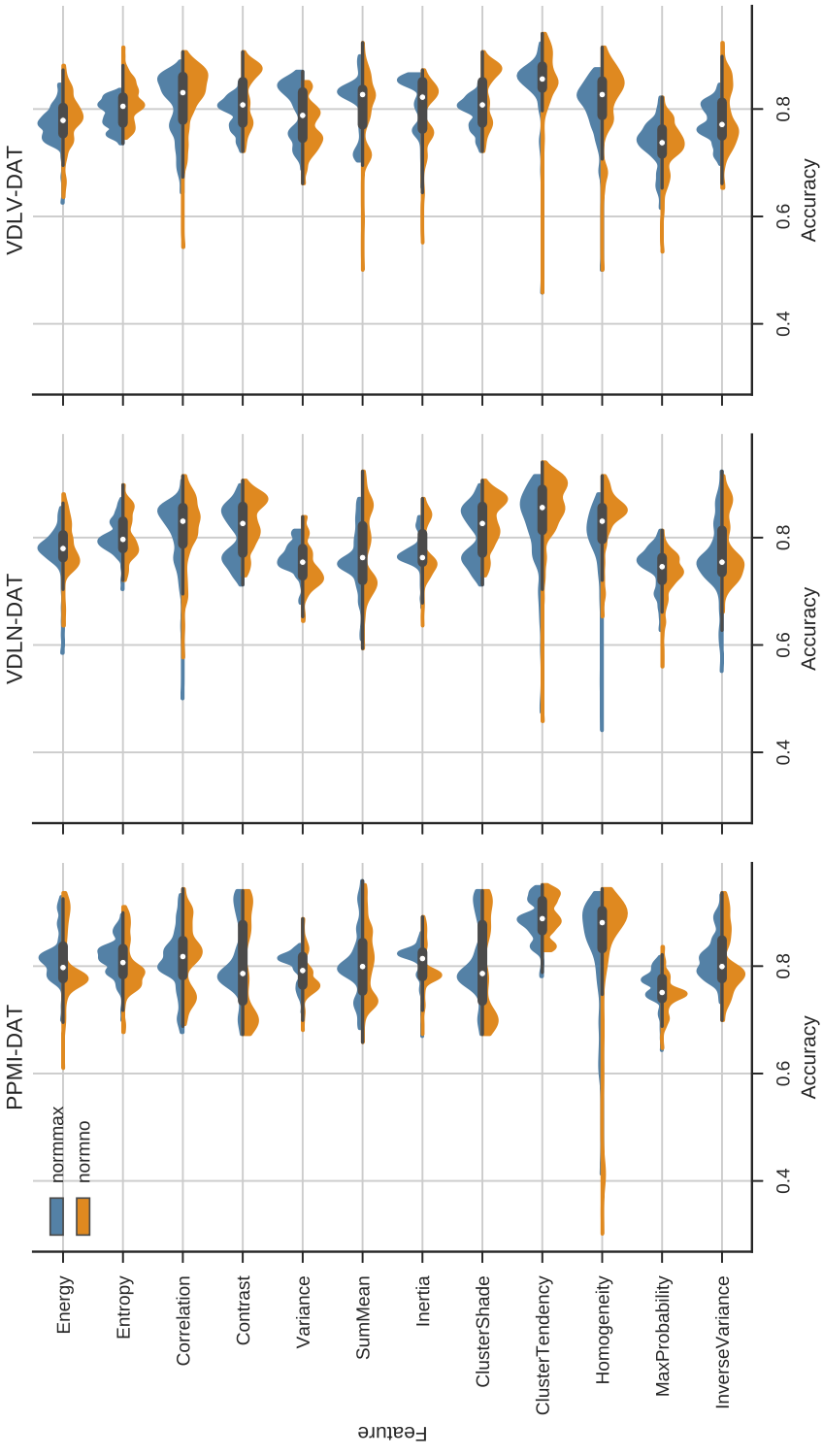
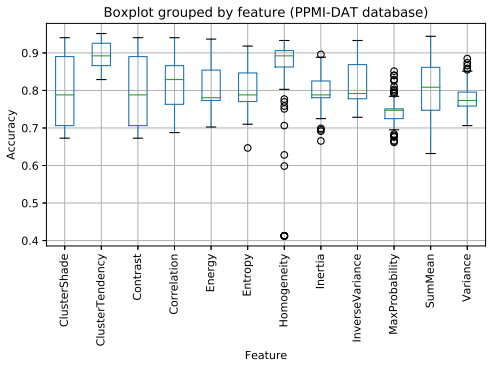
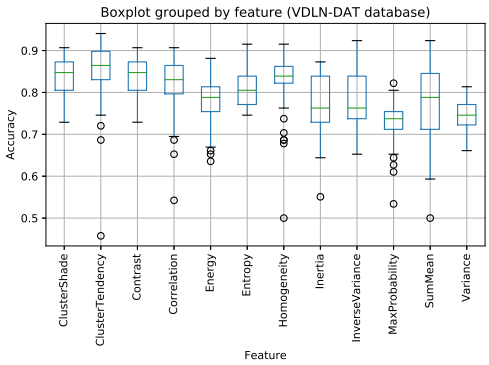


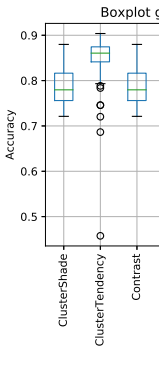
Figure 5.4: Violin plot of all accuracy values, grouped by database and showing the differences between normalization to the maximum and the original images.



(a) PPMI database



(b) VDLN database



(c

Figure 5.5: Box plot of all 130 accuracy values computed for each feature, using the "single approach", at 10 distances d (ranging from 1 to 10) and 13 spatial directions, for (a) PPMI database, (c) VDLV database and (b) VDLN database. The red marks represent the outliers.

single approach also performs relatively well, proving the value of the Haralick texture features to characterize DaTSCAN images.

Results are particularly good in every case when $I_{th} > 0.30 \times I_{max}$, a phenomenon that was previously shown in PPMI database (see Fig. 5.3), but that also extends here to all other databases.

5.3.2 Experiment 2

For experiment 2, all features computed in the aforementioned experiments (the 13 direction vectors and 10 distances used to compute the 3D coocurrence matrix, and the 12 Haralick texture features extracted from these matrices) are used as an input to the classifier. But, in order to reduce dimensionality, we use the measures of discrimination ability proposed in Section **??** to rank these features in a descending order of ability in distinguish PD patterns from normal controls, selecting the first N.

Firstly, the impact of our volume selection threshold I_{th} (see Sec. 5.2.1) on the quality of the resulting Haralick Features, and thus, the accuracy of the experiment, will be analysed. As commented before, best results should be obtained when taking into account the biggest volume of the brain containing only brain voxels, and thus, eliminating the background. Regarding all databases, we obtain Figure 5.6, in which average values of accuracy for every value of I_{th} are plotted. These average values are computed in a similar way to Fig. 5.3, by aver-

aging all 50 accuracy values that correspond to each value of I_{th}. The accuracy values are obtained using each of the 5 proposed selection methods, and using each value of percentage of features selected (ranging from 1 to 100%, by steps of 10%, of the total amount of 1560 features).

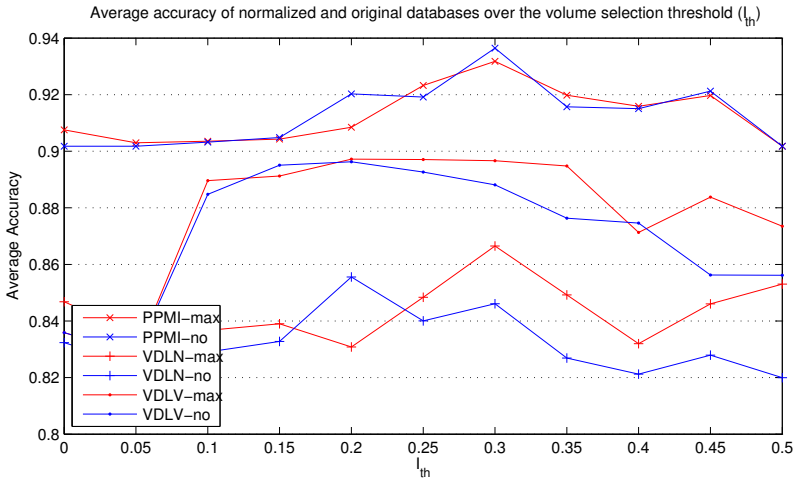
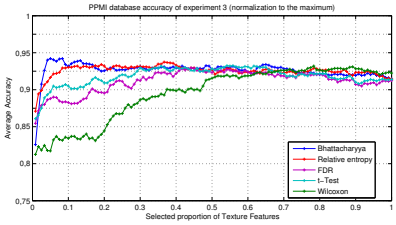


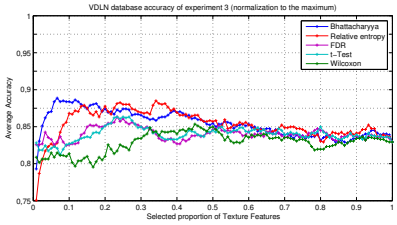
Figure 5.6: Accuracy obtained by averaging all accuracy values using a given volume selection threshold I_{th}

For two out of three databases there is a clear maximum in accuracy for an $I_{th} = 0.30 \times I_{max}$, while the remaining one obtain similar results along a wide range of I_{th}. Furthermore, best average values are obtained using the normalized database, although the PPMI case is slightly different, due to the attenuation correction preprocessing. In Fig. **??** the resulting subimage of applying this threshold was shown, to provide a better understanding of how the textural features are better defined in this. As suggested before, all no-brain voxels are removed from this images, all the textural features correspond only to the internal brain textural changes, and thus, to the textural patterns of the disease, leading to a better performance.

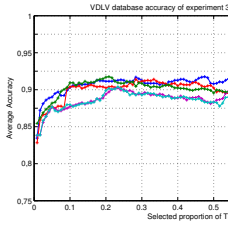
As results suggest, the use of our volume selection strategy with a intensity threshold between 0.25 and 0.45 is profitable in all cases. Also, the use of intensity normalized images, using the normalization to the maximum algorithm has also a good impact on the performance of the system. In this context, Fig. 5.7 analyses the behavior of our system using each of the discrimination-based ranking methods. On these three figures, the values of the computed average accuracy (using the values for intensity thresholds of 0.10 to 0.45) are plotted over the percentage of selected features (previously ranked from the most to the least discriminant, following different criteria) using the three databases.



(a) PPMI database



(b) VDLN database



(c) VDLV da

Figure 5.7: Average accuracy computed for each selection criteria, using all accuracy values for intensity thresholds of 0.10 to 0.45. These values are plotted over N, the number of features selected using some of the ranking criteria defined in Sec. **??** (where N ranges from 1% and 100% of the 1560 total Haralick features calculated). These values correspond to the images of the (a) PPMI database, (c) VDLV database and (b) VDLN database (experiment 2).

Some conclusions about the amount of features that each of our discrimination-ranking methods need can be extracted from these average accuracy graphs. Methods that obtain their maximum accuracy using less than 50% of the features can be considered of great help, as they perform a significant feature reduction. Therefore, methods like Mann-Whitney-Wilcoxon (MWW) can no longer be considered, as it needs more than 50% of features to obtain good results. The opposite behaviour is given by Bhattacharyya Distance (BD) and Relative Entropy (RE), that obtain their maximum average accuracy using the first 10% of features. Fisher's Discriminant Ratio (FDR) and t-Test need a higher amount, but less than 50%.

The aforementioned behaviour correspond to an average behaviour in accuracy. To take a deeper look at the different evaluation parameters and different selection criteria, peak results obtained with each selection criteria are shown on Table 5.2.

This table confirms that the Mann-Whitney-Wilcoxon method can be no longer considered, as it needs more than a 50% of the features to obtain poorer results than all others. Regarding the remaining methods, we observe that those that needed a lower amount of features (BD and RE) obtain here lower values of accuracy that those that needed a higher amount (t-Test and FDR). So, the choice of the best method is, in this case, a matter of trade-off between the computer performance (the number of features to estimate) and the accuracy needed. As in clinical practice, accuracy (and PL) is the parameter that needs to be maximized, we can conclude that FDR and t-Test are the best discrimination-ranking methods to use in this task, although all other methods reveal the ability of our

Database	Criterium	Accuracy	Sensitivity	Specificity	PL	NL	%
PPMI	Bhattacharyya	0.967	0.973	0.962	25.62	0.028	49.1
	Relative Entropy	0.967	0.982	0.956	22.16	0.019	30.8
	FDR	0.974	0.991	0.962	26.10	0.009	34.2
	t-Test	0.974	0.991	0.962	26.10	0.009	35.8
	Wilcoxon	0.959	0.955	0.962	25.15	0.047	85.8
VDLN	Bhattacharyya	0.924	0.956	0.904	9.97	0.049	10.0
	Relative Entropy	0.924	0.933	0.918	11.36	0.073	20.0
	FDR	0.941	0.933	0.945	17.03	0.071	16.7
	t-Test	0.932	0.933	0.932	13.63	0.072	22.5
	Wilcoxon	0.898	0.889	0.904	9.27	0.123	3.3
VDLV	Bhattacharyya	0.938	0.935	0.940	15.59	0.069	40.8
	Relative Entropy	0.933	0.935	0.930	13.36	0.070	45.8
	FDR	0.928	0.963	0.890	8.75	0.042	30.8
	t-Test	0.933	0.935	0.930	13.36	0.070	34.2
	Wilcoxon	0.928	0.926	0.930	13.23	0.080	18.3

Table 5.2: Best results obtained in experiment 2, using three databases, in terms of its accuracy, sensitivity, specificity, Positive Likelihood and Negative Likelihood. The amount of features used to achieve these results is shown as a percentage of the total number of features (1560). Values obtained by leave-one-out.

system in the PD detection with an relevant performance (over 90% of accuracy in most cases).

For comparison purposes, we have established a baseline method proposed in Illan et al [**Illan2012**], where a Voxels-as-Features (VAF) approach with SVM linear, using different normalization strategies were tested. Two additional methods have been compared with our proposed system in order to check the performance versus state-of-the-art algorithms. These methods have been an asymmetrical Single Value Decomposition (SVD) [**Segovia2012**] that appplied SVD on both sides of the brain (since PD often appears only in one hemisphere), and a Empirical Mode Decomposition (EMD) [**Rojas2012**] using different Independent Mode Functions (IMF), particularly the IMF-3. Table 5.3 compares all the aforementioned methods.

System	Acc	Sens	Spec	PL	NL
Homogeneity	0.959	0.973	0.949	19.22	0.028
Cluster Shade	0.955	0.964	0.949	19.01	0.038
Cluster Tendency	0.955	0.973	0.943	17.10	0.029
Correlation	0.941	0.946	0.937	14.92	0.058
Energy	0.937	0.964	0.918	11.73	0.039
Entropy	0.967	0.982	0.956	22.16	0.019
FDR	0.974	0.991	0.962	26.10	0.009
t-Test	0.974	0.991	0.962	26.10	0.009
VAF	0.840	0.807	0.862	5.88	0.224
VAF-IN	0,913	0.890	0.932	13.08	0.118
SVD	0.940	0.962	0.918	11.73	0.041
EMD-IMF3	0.950	0.951	0.948	18.28	0.051

Table 5.3: Comparison of our proposed system (using different texture features) and some other methods in the bibliography: VAF system using the intensity-normalized images, a combination of intensity normalization strategies and classifiers (VAF-IN) [**Illan2012**], a SVD-based approach [**Segovia2012**] and EMD using the third independent mode function (IMF3) [**Rojas2012**].

In Table 5.3, performance values at the operation point are shown for Experiment 1 (using five texture features: Homogeneity, Cluster shade, Cluster tendency, Correlation and energy) and for Experiment 2 (using Relative Entropy, FDR or Student's t-Test). These values are compared with the VAF, SVD and EMD approaches previously cited. We can observe that the performance values obtained with Experiment 1 are very similar to other state-of-the-art methods, like the proposed in [**Segovia2012**, **Rojas2012**], whereas the methodology used

in Experiment 2 outperform all previously used methods. Particularly, as we previously mentioned, the use of either FDR or t-Test to select the most discriminant features gives us results over a PL of 26 and sensitivity over 99%, which proves the ability of some Haralick Textures, and the combination of them, in characterizing the different Parkinson's Disease patterns, and the robustness of the proposed methods.

5.4 Discussion

As discussed later in this Section, the optimum sub-volume will have a smaller size than $40 \times 40 \times 50$, and so, the maximum value of $d = 10$ correspond to at least 20% of the brain sub-volume selected at that point; lower frequency textural changes can be neglected for diagnosis. Furthermore, as the voxel size of all databases is approximately 2x2x2mm, the maximum textural changes are computed within a 20x20x20mm area. This is approximately half the size of the striatum, which is enough to correctly extract the textural features of the area.

The optimum value of I_{th} should be high enough to avoid introducing background voxels in the subvolume selected, yet adequately low to select the biggest subvolume containing only brain voxels. This should lead to the best performance, since the 3D GLC matrices (and the Haralick Texture Features) would have enough information, and would not include non-brain textural patterns.

From these graphs it is patent that normalization has no impact on the performance achieved by our system, which can be consider an advantage, since it reduces the preprocessing needed.

In this regard, the VDLV-DAT dataset has an strange behaviour. It performance holds and even increases when using normalization to the maximum, but significantly drops when integral normalization is used. This is exactly the opposite as happens to the other dataset, and will be discussed later.

6 SPHERICAL BRAIN MAPPING

6.1 Introduction

The main aim of this work is to provide a new framework that allows the mapping of a 3D brain image to a two-dimensional space by means of some statistical measures [**Martinez-Murcia2015**]. The system is based on a conversion from 3D spherical to 2D rectangular coordinates. For each spherical coordinate pair (θ, φ), a vector containing all voxels in the radius is selected, and a number of values are computed, including statistical values (average, entropy, kurtosis) and morphological values (tissue thickness, distance to the central point, number of non-zero blocks). These values conform a two-dimensional image that can be computationally or even visually analysed. In this paper, we proceed using an statistical mask computed using a two-sample t-Test, and a SVM classifier. We have achieved performance results that match those obtained with a priori information but using a complete automated procedure, which furthermore reduces the dimensionality of the original images from more than two million voxels to hardly tens of thousands pixels. These maps can be successfully used by itself, but also allow further processing by using them combined and applying some univariate or multivariate algorithms for dimensional reduction.

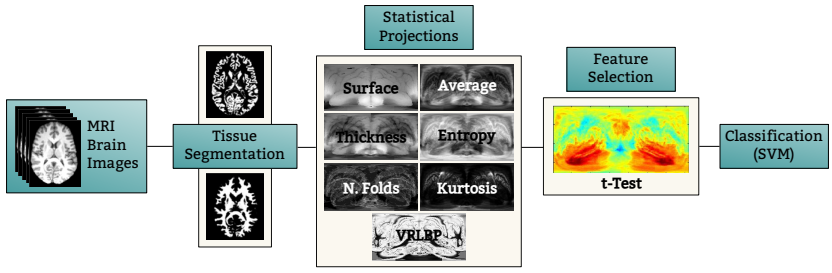


Figure 6.1: Flow diagram of the procedure used in the textural analysis of projected MR brain images.

In [**Martinez-Murcia2015**] a new framework called Spherical Brain Mapping (SBM) was proposed. It was intended to perform feature extraction in MR Brain Images by reducing the whole images to bidimensional maps representing a

number a number of statistical and morphological measures. Each pixel in the maps was the result of computing a certain measure on a set of voxels crossed by the mapping vector, a rectilinear vector centred at the Anterior Commissure (AC) and spanning over all values of azimuth and elevation angles. The bidimensional maps were related to anatomical changes such as brain atrophy or cortical thickness, yielding high performance in differential diagnosis. Furthermore, they provided a significant feature reduction, as well as a visual representation of the underlying information. An extension to the framework using texture and volumetric Local Binary Patterns (LBP) [**Unay2007**] around the mapping vector was proposed in [**Martinez-MurciaVRLBP**].

In this work, we propose a new path tracing algorithm, based on Hidden Markov Models (HMMs), to enhance the mapping procedure in SBM by replacing the mapping vector with curvilinear paths that adapt to the structural information present in MRI. This allows the computation of the feature maps as well as the direct use of the intensity distribution along the path as a characterization of the structural differences in normal or AD-affected subjects. Since the Grey Level Co-occurrence Matrix (GLCM), firstly developed by Haralick[**Haralick73**], have been used in the characterization of brain in numerous works[**martinez2014parametrization sikio2015mr**], we propose a possible extension intended to characterize the brain texture along each path and its neighbourhood.

6.2 Spherical Brain Mapping

Original SBM [**Martinez-Murcia2014225**, **Martinez-MurciaVRLBP**, **Martinez-Murcia2015**, **Martinez-Murcia2016**]

The technique proposed to perform the mapping of the 3D brain images to a 2D map using spherical coordinates -from now on Spherical Brain Mapping (SBM)- is based on the use of spherical coordinates in the brain. A base point is set in the central voxel of the MRI image, and a mapping vector $\mathbf{v}_{\theta,\varphi}$ of length N is defined for each inclination (θ) and azimuth (φ) angles in the range $0° < \theta < 180°$ and $0° < \varphi < 360°$ (see Figure 6.2). Therefore, we will define the sampled set $V_{\theta,\varphi}$, a set that contains P voxels crossed by the sampling vector $\mathbf{v}_{\theta,\varphi}$. For each set $\mathbf{v}_{\theta,\varphi}$, a mapping value ν is computed from the sampled voxels $V_{\theta,\varphi}$, depending on the measure used. In this section, six basic measures are proposed:

- A basic brain surface approach, that accounts for the distance between the central voxel and the last tissue voxel in $V_{\theta,\varphi}$ that is greater than a thresh-

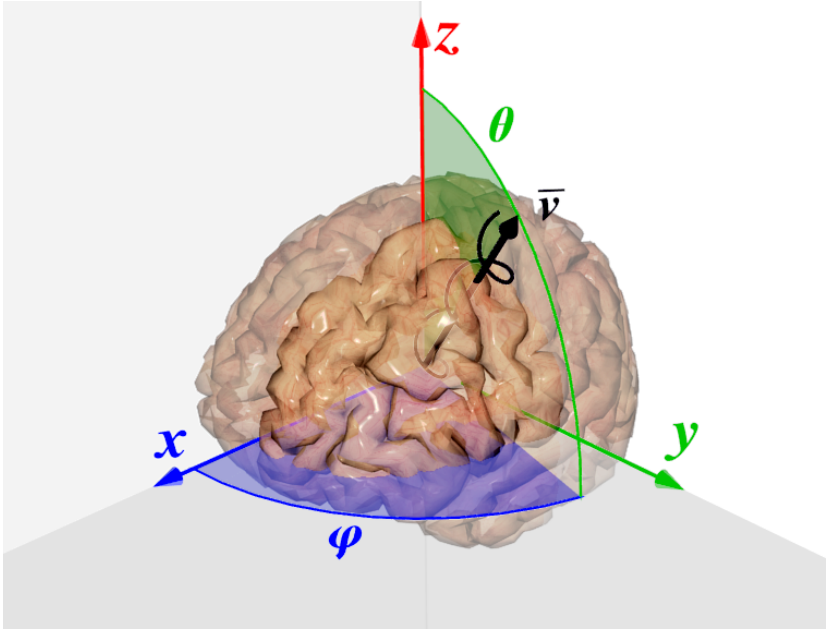


Figure 6.2: Illustration of the computation of the mapping vector $\mathbf{v}_{\theta,\varphi}$, the angles θ and φ and the r-neighbourhood of $\mathbf{v}$ (see Section 6.3).

old I_{th}. This might allow our system to observe structural degeneration and tissue loss in the surface of the tissue.

$$v_{surf} = \arg\max_{i}\{V_{\theta,\varphi}(i) > I_{th}\} \quad \forall i = 1,\dots P \tag{6.1}$$

- Another parameter used is thickness of the tissue. This can be useful when measuring the thickness of segmented Gray Matter or White Matter MR images. It is defined as the distance between the last and first elements in $V_{\theta,\varphi}$ with an intensity greater than a threshold I_{th} (typically 0):

$$v_{thick} = \arg\max_{i}\{V_{\theta,\varphi}(i) > I_{th}\} - \arg\min_{i}\{V_{\theta,\varphi}(i) > I_{th}\} \quad \forall i = 1,\dots P \tag{6.2}$$

- The number of folds represents the number of overlapping segments of tissue in the set $V_{\theta,\varphi}$. It is computed by thresholding $V_{\theta,\varphi}$ using the value I_{th} and counting the number of resulting connected subsets. Let $A_{\theta,\varphi}$ be the set that contains all the indices of the voxels in $V_{\theta,\varphi}$ with an intensity greater than I_{th}:

$$A_{\theta,\varphi} = \{i \,/\, V_{\theta,\varphi}(i) > I_{th}\} \tag{6.3}$$

where $A_{\theta,\varphi} \in \mathbb{N}$. Let us divide $A_{\theta,\varphi}$ in J disjoint connected subsets so that:

$$A_{\theta,\varphi} = A^1_{\theta,\varphi} \cup A^2_{\theta,\varphi} \cup \cdots \cup A^J_{\theta,\varphi} \quad \text{so that} \quad A^i_{\theta,\varphi} \cap A^j_{\theta,\varphi} = \emptyset \quad \forall i,j \tag{6.4}$$

Therefore, our $v_{nf} = J$, the number of disjoint connected subsets in $A_{\theta,\varphi}$.

- An average approach, where the average of all the intensity values in the set $V_{\theta,\varphi}$ is computed as:

$$v_{av} = \frac{1}{N} \sum_i V_{\theta,\varphi}(i) \quad \forall i = 1, \ldots P \tag{6.5}$$

- The entropy assumes that the set $V_{\theta,\varphi}$ is a probability mass vector (probability of belonging to a certain tissue, normalized) and computes v as:

$$v_{ent} = \sum_i V_{\theta,\varphi}(i) * \log(V_{\theta,\varphi}(i)) \quad \forall i \in \arg_i\{V_{\theta,\varphi}(i) > 0\} \tag{6.6}$$

- The uncorrected kurtosis, also known as fourth standardized moment, of the set $V_{\theta,\varphi}$ in which v is calculated using:

$$v_{kurt} = \frac{\frac{1}{N}\sum_i \left(V_{\theta,\varphi}(i) - \bar{V}_{\theta,\varphi}(i)\right)^4}{\left(\frac{1}{N}\sum_i \left(V_{\theta,\varphi(i)} - \bar{V}_{\theta,\varphi}(i)\right)^2\right)^2} \quad \forall i = 1, \ldots P \tag{6.7}$$

where $\bar{V}_{\theta,\varphi}$ is the average of all voxels in $V_{\theta,\varphi}$ (same value as v_{av}, described in Eq. 6.5).

The resulting GM and WM maps are depicted in Figure 6.3.

The resulting maps will contain the value v mapped in each direction (θ, φ). As the inclination angle θ ranges from 0° to 180° and the azimuth φ from 0° to 360°, the resulting maps will have a size of 181×361, where each pixel is the v value for each direction. The whole algorithm can be downloaded at http://wdb.ugr.es/~fjesusmartinez/portfolio/sbm/.

This methodology defines the sampling set as the voxels that are crossed by the sampling vector $\mathbf{v}_{\theta,\varphi}$. When projecting a structure as complex as the brain, this implies a loss of contextual information of both the neighbourhood and the layers crossed by $\mathbf{v}_{\theta,\varphi}$. Two approaches have been suggested to overcome this problem: the first one extends the system by dividing the sampling set $V_{\theta,\varphi}$ in n equal parts in a so-called "Layered approach", and the second one uses a helical sampling and Local Binary Patterns (LBP) to map the neighbourhood of the sampling vector and characterize the texture of the area.

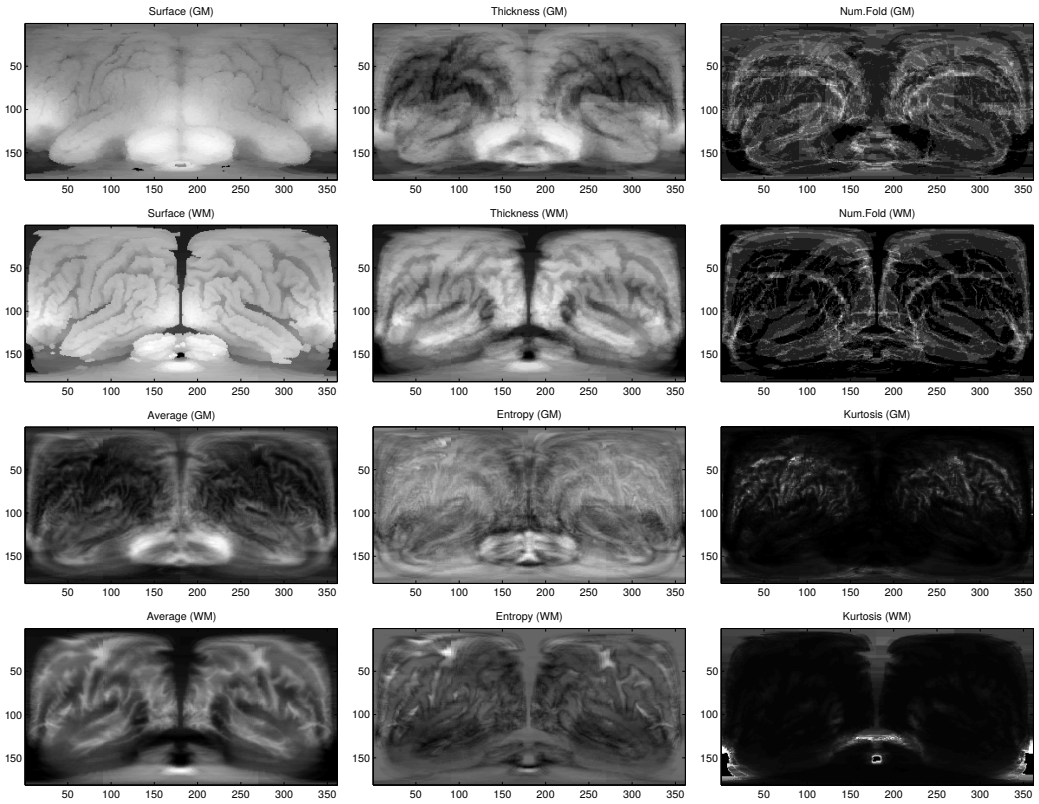


Figure 6.3: Resulting GM and WM maps of the same control subject using the six proposed measures: Surface, Thickness, Number of Folds, Average, Entropy and Kurtosis.

6.2.1 Layered Extension

The first strategy used to improve the descriptive abilities of the mappings is the Layered Extension. In this approach, the sampled voxels set $V_{\theta,\varphi}$ can be divided in n equal subsets, using each one to project one section -or layer- of the brain. If, for example, we use $n = 4$, 4 subsets containing the same number of voxels will be derived from $V_{\theta,\varphi}$, and therefore 4 different maps will be depicted, from the closest to the centre to the farthest. This layered approach reveals the different anatomical structures found in the brain at different depths, potentially revealing a more detailed distribution the features of the mappings . These layered maps and their performance will be discussed later.

6.3 Volumetric Radial LBP

In a second attempt to overcome the loss of per layer information, we have expanded the influence of $\mathbf{v}_{\theta,\varphi}$ to its r-neighbourhood. This has been done by defining a ν measure that describes not only the features of the voxels crossed by $\mathbf{v}_{\theta,\varphi}$, but the texture of its neighbourhood using a Volumetric Radial Volumetric Radial Local Binary Pattern (VRLBP) descriptor.

Local Binary Patterns (LBP) were first introduced in **[Ojala1996]** to describe the texture of an image with application to face recognition. Later, in **[Zha02007]**, the technique was extended to a Volume LBP (VLBP), defining a 3D texture in a local neighbourhood by using a cylinder oriented in one direction and whose radius define the neighbourhood used to compute the LBP descriptor.

In the VRLBP, the sampling method devised in **[Zha02007]** has been updated to follow helical coordinates around the mapping vector $\mathbf{v}_{\theta,\varphi}$ (see helix around $\mathbf{v}_{\theta,\varphi}$ in Figure 6.2). Formally, we note $V^{P,r}_{\theta,\varphi}$ the set of P sampled voxels of the image I in the r-neighbourhood of $\mathbf{v}_{\theta,\varphi}$ taken by helical sampling:

$$V^{P,r}_{\theta,\varphi} = \{I(g^{0,r}_{\theta,\varphi}), I(g^{1,r}_{\theta,\varphi}), I(g^{2,r}_{\theta,\varphi}), \dots I(g^{P-1,r}_{\theta,\varphi})\} \tag{6.8}$$

where the coordinates $g^{p,r}_{\theta,\varphi}$ of each voxel are computed in the direction of $\mathbf{v}_{\theta,\varphi}$ by:

$$g^{p,r}_{\theta,\varphi} = \begin{cases} x^{p,r}_{\theta,\varphi} = p\sin(\varphi)\cos(\theta) - r\sin(2\pi n\frac{p}{P}) \\ y^{p,r}_{\theta,\varphi} = p\sin(\varphi)\sin(\theta) + r\cos(2\pi n\frac{p}{P}) \\ z^{p,r}_{\theta,\varphi} = p\cos(\varphi) \end{cases} \qquad p = \{0, ..., P-1\}, P \in \mathbb{N} \tag{6.9}$$

being n the number of loops in the helix. Following [**Zhao2007**], voxels that do not fall exactly at the coordinates computed in the equations 6.9 are estimated by interpolation.

Let us assume, without lost of generality, that P and r are fixed. This way, set of sampled voxels $V^{P,r}_{\theta,\varphi}$ becomes $V_{\theta,\varphi}$, which matches the definition of Section **??**. Following this notation, the value v for this VRLBP approach is computed using:

$$v_{VRLBP} = \sum_{i} s(V_{\theta,\varphi}(i) - V_{\theta,\varphi}(0)) \cdot 2^i \quad \forall i = 1, \dots P \tag{6.10}$$

where $s(x)$ is the sign function, defined as:

$$s(x) = \begin{cases} 1 & x \geqslant 0 \\ 0 & x < 0 \end{cases} \tag{6.11}$$

This approach provides textural information about all brain structures in a certain direction, as it can be seen in Figure 6.4.

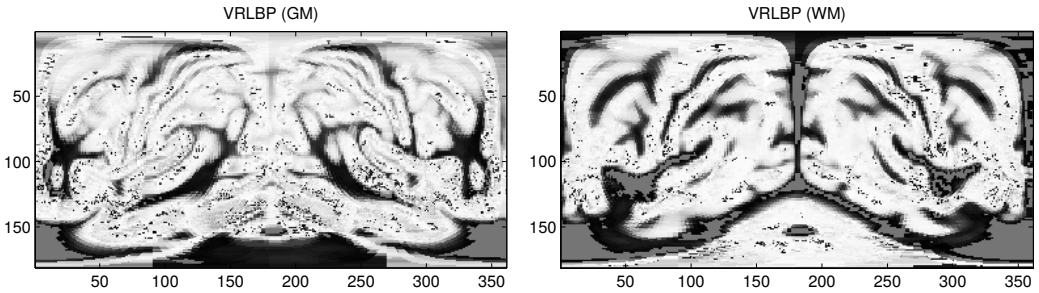


Figure 6.4: An example of the VRLBP projection for GM and WM Tissues.

```
def foo():
      hola amigo
      print('amigo')

eh = foo("amigo")

string title = "This is a Unicode $\pi$ in the sky"
/*
Defined as $\pi=\lim_{n\to\infty}\frac{P_n}{d}$ where $P$ is the
    perimeter
of an $n$-sided regular polygon circumscribing a
circle of diameter $d$.
*/
const double pi = 3.1415926535
```

6.4 Path via Hidden Markov Models

We have already talked about the limitations of the rectilinear mapping vector used in [**Martinez-Murcia2015**] and [**Martinez-MurciaVRLBP**]. Therefore, it would be desirable to define new mapping paths that use all the available intensity and spatial information in MRI images. This way, the resulting sets of selected voxels (and the resulting maps) would contain information about both the intensities and structure.

To define the paths, we can define 3D Biomedical Images as a tuple containing spatial information in the image range ($\mathbf{x} \in \mathbb{I}$, where $\mathbb{I} \subset \mathbb{R}^3$) as well as intensity information ($I(\mathbf{x}) \in \mathbb{R}$). There exist a number of possibilities in the interpretation of intensity data on the images, from plain intensity values to a sampling of the underlying tissue density (and thus, an estimation of the probability of finding tissue in each position).

Following our SBM approach, in which some radia are used to extract relevant statistical features from these images, we formulate a 3D path tracing algorithm suitable for extraction of curvilinear structures from 3D biomedical images, and directly linked to each direction (φ, θ) as in the original work. Our objective is to make these paths as representative of the underlying intensity distribution as possible. To do so, we must use both intensity and spatial information to construct maximum intensity-similarity paths oriented in the given direction.

Let us note a 3D path in a certain direction (φ, θ) as a Markov Model[**Chen2008**]:

$$\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_N\} \tag{6.12}$$

Therefore, our optimum path would be the one that maximizes the probability of the path:

$$\mathbf{X}_{opt} = \arg\max_{\mathbf{X}} \{P(\mathbf{X})\} \tag{6.13}$$

or, similarly, the probability of all the nodes:

$$P(\mathbf{X}) = P(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_N) \tag{6.14}$$

where $\mathbf{x}_0$ is the node located at the AC and $\mathbf{x}_N$ is the theoretical projection of the current direction (φ, θ) in the limits of the image $\mathbb{I}$. Setting $\mathbf{x}_0$ at the AC is not a random choice; it is a convention when using the MNI coordinates[**Evans1993**] and furthermore, since it is a point that shares connectivity with both hemispheres, the resulting paths will be optimal, covering most of the brain. We can assume a first-order Hidden Markov Model (HMM) for the tracing of the path, and the i-th node will be computed as:

$$P(\mathbf{x}_i | \mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \dots \mathbf{x}_0) \approx P(\mathbf{x}_i | \mathbf{x}_{i-1}) \tag{6.15}$$

and with this assumption, (6.14) becomes:

$$P(\mathbf{X}) = P(\mathbf{x}_0, \mathbf{x}_1, \dots \mathbf{x}_N) = \prod_{i=1}^{N} P(\mathbf{x}_i | \mathbf{x}_{i-1}) \tag{6.16}$$

In our case, we will assume that the hidden state of each node will be its intensity $I(\mathbf{x}_i)$. Let $\mathbf{I} = \{I(\mathbf{x}_0), I(\mathbf{x}_1), \dots I(\mathbf{x}_N)\}$ be the vector of the intensities in each node. With the introduction of these factors, our optimal path defined in (6.13) can be view as:

$$\mathbf{X}_{opt} = \arg\max_{\mathbf{X}} \{P(\mathbf{X}|\mathbf{I})\} \tag{6.17}$$

$$P(\mathbf{X}|\mathbf{I}) = P(\mathbf{x}_0, \dots x_N | I(\mathbf{x}_0), \dots I(\mathbf{x}_N)) \tag{6.18}$$

$$= \frac{P(I(\mathbf{x}_0), \dots I(\mathbf{x}_N)|\mathbf{x}_0, \dots x_N) \cdot P(\mathbf{x}_0, \dots x_N)}{P(I(\mathbf{x}_0), \dots I(\mathbf{x}_N))} \tag{6.19}$$

where:

$$P(I(\mathbf{x}_0), \dots I(\mathbf{x}_N)|\mathbf{x}_0, \dots x_N) = \prod_{i=1}^{N} P(I(\mathbf{x}_i)|\mathbf{x}_i) \tag{6.20}$$

and $P(I(\mathbf{x}_0), \dots I(\mathbf{x}_N))$ is the *a priori* probability of the intensities in the path. We can assume, without lack of generality, that this term is constant along the path, and therefore, it plays no part in the optimization process.

For computational purposes, we will compute all the needed probabilities on a set of candidates $\mathbf{X}_c = \{\mathbf{x}_{c,1}, \mathbf{x}_{c,2}, \dots \mathbf{x}_{c,M}\}$ defined by $\mathbf{x}_{i-1}$. This candidates are contained inside the L^2-norm support ball $B_{2,r}(\mathbf{x} - \mathbf{x}_{i-1})$ of radius r centred in $\mathbf{x}_{i-1}$.

To estimate the individual probabilities needed for (6.20) $P(I(\mathbf{x}_i)|\mathbf{x}_i)$, we can assume a normal distribution of intensities of the candidates with mean $I(\mathbf{x}_{i-1})$ and variance σ_c^2 the variance of the intensities of the candidate set. By assuming this, the probability of a certain intensity in the candidate node $\mathbf{x}_i$ increases as the intensity becomes more similar to the intensity of $\mathbf{x}_{i-1}$. Therefore, the assumption supports the tracing of minimal intensity difference paths. The probability of the intensity of a candidate $\mathbf{x}_i$ will be:

$$P(I(\mathbf{x}_i)|\mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(I(\mathbf{x}_i) - I(\mathbf{x}_{i-1}))^2}{2\sigma_c^2}\right) \tag{6.21}$$

The last term $P(\mathbf{x}_0, \dots \mathbf{x}_N)$ to be defined in (6.19) is directly related to the SBM framework defined before, as it will depend on the radial direction (φ, θ) that we want to "force" in the path. To define this term, we will define an attractor located in the position $\mathbf{x}_N$ (the projection of the current direction (φ, θ) in

the limits of the image $\mathbb{I}$). We assume that this attractor defines the conditionality, that is, that it affects the transition probability between states by means of an isotropic Gaussian radial basis function (RBF), as defined in (6.22). This definition helps the attractor to lightly condition the direction of the path at first, and more strongly as the path approaches the cortex, leading to a better representation of the underlying structure.

$$P(\mathbf{x}_0, \dots \mathbf{x}_N) = P(\mathbf{x}_i|\mathbf{x}_N) \tag{6.22}$$

$$= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_N)\Sigma^{-1}(\mathbf{x}_i - \mathbf{x}_N)\right) \tag{6.23}$$

where Σ is the covariance matrix of the given distribution. As we will consider only isotropic gaussian kernel, the matrix can be considered as a 3×3 diagonal matrix whose diagonal elements are a scalar value σ^2, which we set in each iteration to the euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_N$.

6.4.0.1 *Step Size*

In this algorithm, instead of using a fixed step size, we evaluate each candidate point $\mathbf{x} \in B_{2,r}(\mathbf{x} - \mathbf{x}_i)$, and thus, the only parameter to regulate is the radius of the L2-norm support ball r. As a trade-off between computational issues and definition of the defined path, we will use a value of $r = 3$ voxels, which gives approximately 200 candidate points in each iteration.

6.4.0.2 *Stop Condition*

Although the paths could be defined until they reach the last point $\mathbf{x}_N$ (remember, the projection of the general direction onto the limits of $\mathbb{I}$), our interest is to model the paths inside the brain, and therefore we establish a stop condition using an intensity threshold. This threshold is computed using the entropic thresholding, as defined in **[Yen1995]**. Let $G_m \equiv \{I_0, I_1, ...I_m\}$ denote the set of intensity levels of the whole image. We can compute a histogram to obtain the observed frequencies f_{I_i}, and thus, the observed probability of the different Grey levels $p_i = f_{I_i}/N$, where N is the number of voxels in our image $\mathbb{I}$.

For a given intensity threshold $I_{th} = I_s$, if $\sum_{i=0}^{s-1} p_i$ is larger than zero and smaller than 1, the following distributions can be derived from this distribution after normalization:

$$A \equiv \left\{ \frac{p_0}{P(I_s)}, \frac{p_1}{P(I_s)}, \dots, \frac{p_{s-1}}{P(I_s)} \right\} \tag{6.24}$$

$$B \equiv \left\{ \frac{p_s}{1 - P(I_s)}, \frac{p_{s+1}}{1 - P(I_s)}, \dots, \frac{p_m}{1 - P(I_s)} \right\} \tag{6.25}$$

where $P(I_s) = \sum_i^s p_{I_i}$ is the cumulative density function for the s-th Grey level. Therefore, we choose the threshold so that the total amount of information provided by A and B (foreground and background of the image) is maximized. The total information provided by the s-th Grey level is:

$$TE(s) = E_A(s) + E_B(s) \tag{6.26}$$

$$= -\sum_{i=0}^{s-1} \left(\frac{p_i}{P(I_s)}\right) \log\left(\frac{p_i}{P(I_s)}\right) \tag{6.27}$$

$$-\sum_{i=s}^{m-1} \left(\frac{p_i}{1-P(I_s)}\right) \log\left(\frac{p_i}{1-P(I_s)}\right) \tag{6.28}$$

A summary of our HMM-based path tracing method is shown in Algorithm 1. To illustrate the effect of the algorithm on a real example, Fig. 6.5 depicts the resulting set of paths computed over the standard MRI DARTEL template.

Algorithm 1: HMM-based Path Creation

input : MRI Brain Image I of size $U \times V \times W$, $\mathbf{x}_0$
output: List of nodes in the optimum path $\mathbf{X}_{opt}$

Compute the $I_{th} = I_s$ where s maximizes $TE(s)$;
Set $\mathbf{x}_0$ to the AC;
Compute the attractor position $\mathbf{x}_N$ in the direction (φ, θ);
$\mathbf{x}_i \leftarrow \mathbf{x}_0$;
while $(i < \mathit{IterLimit})$ *&* $(I(\mathbf{x}_i) > I_{th})$ *&* $(\mathbf{x}_i \in \mathbb{I})$ **do**
  Get the node candidates $\mathbf{X}_c = \{\mathbf{x}_{c,1}, \mathbf{x}_{c,2}, \ldots \mathbf{x}_{c,M}\}$ where $\mathbf{x}_{c,m} \in B_{2,r}(\mathbf{x}_{c,m} - \mathbf{x}_i)$;
  Get the intensities of the candidates $I(\mathbf{x}_c) \quad \forall \mathbf{x}_c \in \mathbf{X}_c$;
  foreach $\mathbf{x}_c \in \mathbf{X}_c$ compute $P(\mathbf{x}_c|\mathbf{x}_N)$ and $P(I(\mathbf{x}_c)|\mathbf{x}_i)$;
  $\mathbf{x}_{i+1} = \arg\max_{\mathbf{x}_c}[P(I(\mathbf{x}_c)|\mathbf{x}_i) \cdot P(\mathbf{x}_c|\mathbf{x}_N)]$;
  $i = i + 1$;
$\mathbf{X}_{opt} \leftarrow \{\mathbf{x}_0, \mathbf{x}_1, \ldots \mathbf{x}_N\}$;

6.4.1 Radial Texture Features

The paths extracted with the aforementioned algorithm present a meaningful representation of the structure of the intensity levels (and therefore, the tissue density) of the MRI brain images. At this point, one might consider using the intensity values located in those radia as features, and try to characterize each radius' discrimination ability by means of these intensities.

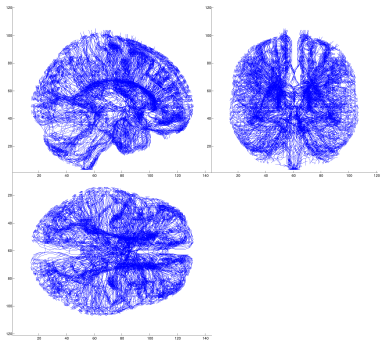


Figure 6.5: Set of HMM based paths over the MRI DARTEL template.

Conversely, the strategy proposed in [**Martinez-MurciaVRLBP**], which uses Local Binary Pattern (LBP) descriptors in the helical neighbourhood of a rectilinear mapping vector, might be a complementary approach. Due to the HMM paths topology, the helical sampling becomes difficult to compute and not even useful, therefore we propose a modification of the Grey Level Co-occurrence Matrix (GLCM) along the paths.

The GLCM, proposed by Haralick[**Haralick73**], is one of the most widely used methods in texture characterization, and it has been successfully applied to medical imaging in the past[**kovalev2001three**, **martinez2014parametrization**]. It works by storing the number of voxel-wise correspondences between K grey levels with a certain position offset Δ on a $K \times K$ matrix (C_Δ) along all the image.

Our modification will compute a node-wise GLC matrix, in which the number of grey-level transitions between adjacent nodes, noted as $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$, is stored along the whole path $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1 \dots \mathbf{x}_N\}$. Mathematically, the computation of the GLCM in each point in the path will be:

$$C_{\Delta_i}(j,k) = \sum_{i=0}^{N-1} \begin{cases} 1 & I(\mathbf{x}_i) = j, I(\mathbf{x}_{i+1}) = k \\ 0 & otherwise \end{cases} \tag{6.29}$$

where the offset is different for each pair of nodes $\Delta_i = \mathbf{x}_{i+1} - \mathbf{x}_i$.

The definition provided in (6.29) is intended for computing the values in each node. However, we can generalize this construction to include not only the nodes, but the intensity information around each node in the computation, which could potentially lead to more significant texture features. Let us note a

set containing all the voxels in the closed neighbourhood of $\mathbf{x}_i$ as X_i. Therefore, (6.29) can be generalized for any voxel $\mathbf{x} \in X_i$ as:

$$C(j,k) = \sum_{i=0}^{N-1} \sum_{\mathbf{x} \in X_i} \begin{cases} 1 & I(\mathbf{x}) = j, I(\mathbf{x} + \Delta_i) = k \\ 0 & otherwise \end{cases} \tag{6.30}$$

From the GLCM, a variety of texture descriptors, or features, can be extracted. In this work we will use ten texture features proposed in the original Haralick's article[**Haralick73**] as well as in Refs [**soh1999texture**] and [**clausi2002analysis**]: Contrast[**Haralick73**], Correlation[**Haralick73**], Dissimilarity[**soh1999texture**], Energy[**Haral**] Entropy[**soh1999texture**], Homogeneity[**Haralick73**], Difference Variance[**Haralick73**] (D. Variance), Difference Entropy[**Haralick73**] (D. Entropy), Inverse Difference Normalized[**clausi2002analysis**] (IDN) and Inverse Difference Moment Normalized[**clausi2**] (IDMN). The mathematical expressions for these features are presented in Equations 6.31 to 6.40.

$$\text{Contrast} = \sum_j \sum_k \{(j-k)^2 C(j,k)\} \tag{6.31}$$

$$\text{Correlation} = \frac{\sum_j \sum_k (j-\mu_j)(k-\mu_k) C(j,k)}{(\sigma_j \sigma_k)} \tag{6.32}$$

$$\text{Dissimilarity} = \sum_j \sum_k \{|j-k| C(j,k)\} \tag{6.33}$$

$$\text{Energy} = \sum_j \sum_k C(j,k)^2 \tag{6.34}$$

$$\text{Entropy} = -\sum_j \sum_k C(j,k) \log(C(j,k)) \tag{6.35}$$

$$\text{Homogeneity} = \sum_j \sum_k \frac{C(j,k)}{1+|j-k|} \tag{6.36}$$

$$\text{D. Variance} = \sum_{j=0}^{N_g-1} j^2 p_{x-y}(j) \tag{6.37}$$

$$\text{D. Entropy} = -\sum_{j=0}^{N_g-1} p_{x-y}(j) \log\{p_{x-y}(j)\} \tag{6.38}$$

$$\text{IDN} = \sum_j \sum_k \frac{C(j,k)}{1+|j-k|/N} \tag{6.39}$$

$$\text{IDMN} = \sum_j \sum_k \frac{C(j,k)}{1+(j-k)^2/N^2} \tag{6.40}$$

6.5 Results

6.5.1 Experimental settings and validation

In this work, have considered a binary classification problem: AD vs. NC, where we have evaluate separately each type of mapping. First, in Section 6.5.2, we will analyse our maps by means of the t statistic, where the areas of higher

statistical significance (AD vs NC) will be highlighted. To better interpret the results, an anatomical reference is provided.

Second, we have perform a classification analysis using feature selection by means of t-Test, then training and testing a linear SVM classifier. The method has been validated using stratified 10-fold cross-validation, as recommended in [**Kohavi1995a**]. This procedure consists on randomly partition the whole datasets into 10 subsets that contains the same proportion of individual of both classes as the whole database. Then, one subset is used for testing and the remaining 9 are used for training. This is repeated for each of the subsets as training sets. Finally, the whole cross-validation strategy will be repeated 10 times to avoid the possible bias and random effects of the partitions, and obtain the average and standard deviation of the performance values.

Values of accuracy (acc), sensitivity (sens) and specificity (spec) along with their standard deviation will be employed to evaluate the performance of the different mappings. Selection of parameter C of the SVM classifier (as implemented in `LIBSVM` [**Chang2001**]) will be performed using an inner 5-fold cross-validation on the training subset.

6.5.2 Statistical Significance Analysis

In this section, we will study the statistical significance of the SBM maps by using a two-sample t-Test with pooled variance estimate, as defined in Section **??**. The computed t values for each coordinate pair in the maps (θ, φ) will be displayed later in Section 6.5.2.2 for the six original measures, in Section 6.5.2.3 for the layered extension and in Section 6.5.2.4 for the VRLBP. However, to provide a better understanding of these t-maps, an anatomical reference is provided in Section 6.5.2.1.

6.5.2.1 *Anatomical Reference*

Our SBM technique maps all sampled voxels selected by our mapping vector $\mathbf{v}_{\theta,\varphi}$ to a single point in the projected map. These points cross different anatomical regions, however it is difficult to know at first which regions are crossed, given the coordinate pairs (θ, φ). To clarify this and provide a better understanding, we have mapped the widely known [**AAL**] atlas [**Tzourio-Mazoyer2002**] using SBM, and the regions are displayed in Figures 6.6 and 6.7.

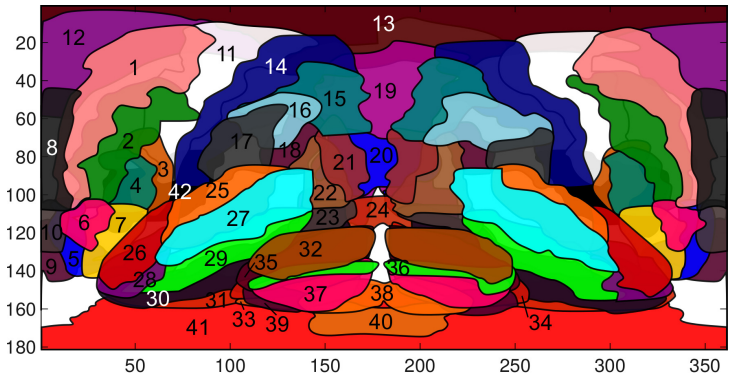


Figure 6.6: Projection of different cortical regions. In the Frontal region, we can find: 1) Frontal Sup., 2) Frontal Mid., 3) Frontal Inf. Oper., 4) Frontal Inf. Tri., 5) Frontal Sup. Orb, 6) Frontal Mid. Orb, 7) Frontal Inf. Orb, 8) Frontal Sup. Medial, 9) Rectus, 10) Frontal Med. Orb., 11) Precentral, 12) Supp. Motor Area. In the Parietal region: 13) Paracentral Lobe, 14) Postcentral, 15) Parietal Sup., 16) Parietal Inf., 17) Supramarginal, 18) Angular. In the Occipital region: 19) Precuneus, 20) Cuneus, 21) Occipital Sup., 22) Occipital Mid., 23) Occipital Inf., 24) Lingual. In the Temporal region: 25) Temporal Sup., 26) Temporal Pole Sup., 27) Temporal Mid., 28) Temporal Pole Mid., 29) Temporal Inf, 30) Fusiform, 31) Parahippocampal. The Cerebellum, divided in: 32) Cerebelum Crus 1, 33) Cerebelum 3, 34) Cerebelum 4-5, 35) Cerebelum 6, 36) Cerebelum 7b, 37) Cerebelum 8, 38) Cerebelum 9, 39) Cerebelum 10. And additionally, the 40) Medulla, 41) Brain Stem and 42) Insula.

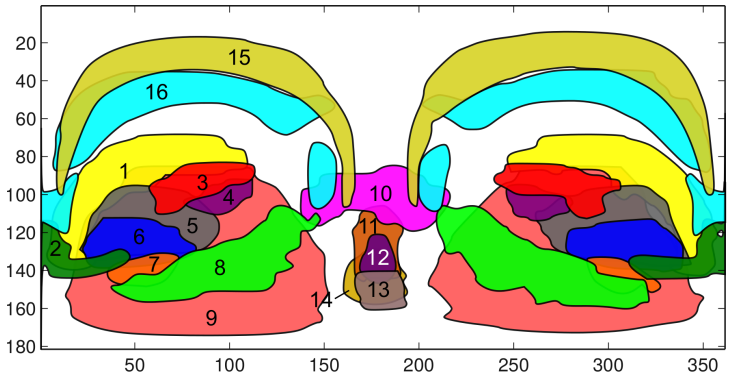


Figure 6.7: Projection of some important subcortical regions and organs. We observe the following subcortical structures: 1) Caudate Nucleus, 2) Olfactory Bulb, 3) Rolandic Operculum, 4) Heschl's gyri, 5) Putamen, 6) Globus Pallidus, 7) Amygdala, 8) Hippocampus, 9) Thalamus, 10) Lingual, 11) Vermis 4-5, 12) Vermis 7, 13) Vermis 9, 14) Vermis 1-2, 15) Cingulate Gyrus, 16) Corpus Callosum

6.5.2.2 *Spherical Brain Mapping*

In this section, the six measures proposed in Section **??** will be analysed using a t-test. To proceed, a t-value will be computed for each pixel in the maps, yielding a significance map. These significance maps, or t-maps, are presented in Figure 6.8.

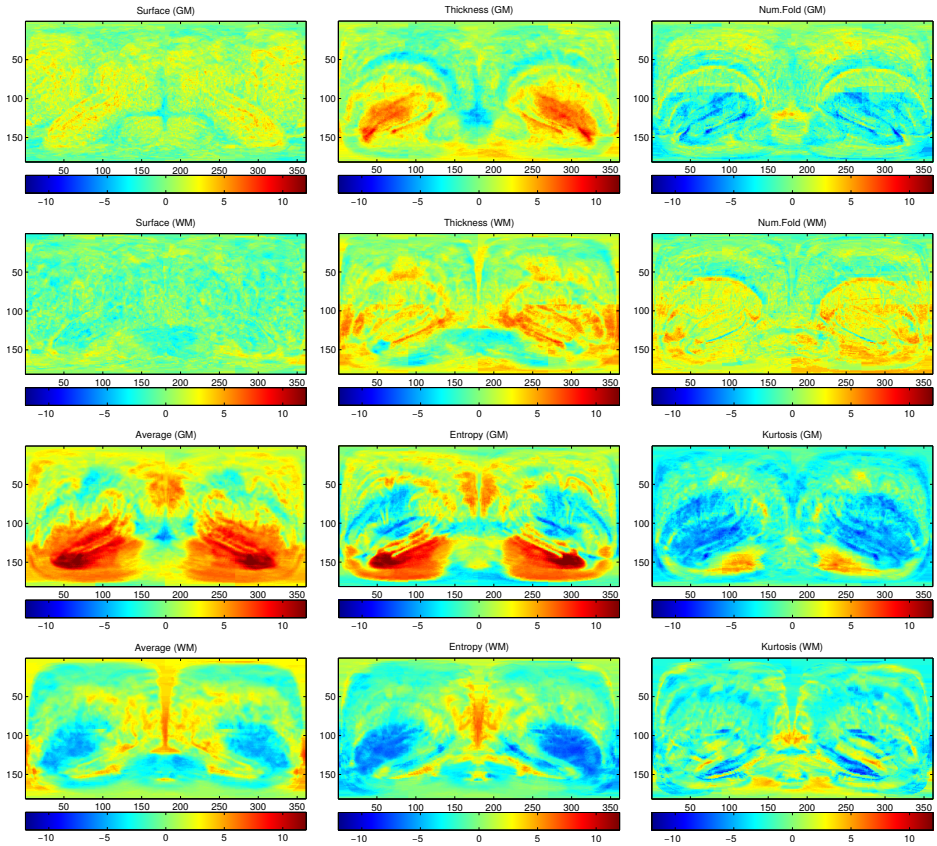


Figure 6.8: t-maps that present the level of statistical relevance in the AD vs. NC paradigm, for each type of mapping and GM and WM.

Absolute t-values higher than 1.96 can be considered to be significant, with a $p < 0.05$. In this case the areas of greater significance are in dark red and dark blue, where red is a positive t-value, meaning a higher value in controls than in AD subjects, and blue is a negative t-value, and conversely, blue means negative values, which are related to a higher value in AD than in controls.

The first thing to note is that the distribution of the t values in the Surface mapping, in both GM and WM, is not relevant at all, with very few significant pixels distributed along the whole image.

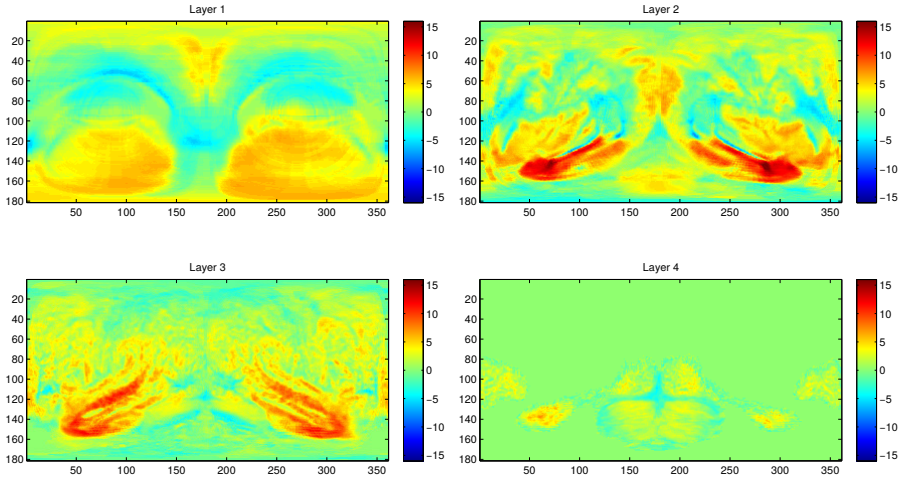
In the remaining GM mappings, greater t-values are located in the frontal, occipital and parietal lobes, but the most significant areas can be found in the temporal lobe. This is most obvious in the Average and Entropy mappings, but can also be found in Thickness. Number of Folds and Kurtosis present high, but however negative, t-values in these areas as well. This suggests that for GM, most of the neurodegeneration is located in the temporal lobe and all the underlying structures that are projected in this area, including the Hippocampus and Parahippocampal gyrus, which are considered a fundamental disease indicator in the NINCDS-ADRDA criteria [**Dubois2007**]. Additionally, some structures that are located in the same area have been recently related to the progression of the disease, such as the Caudate Nucleus and Putamen [**Pievani2013**]. These changes are more precisely located when using one of our spherical maps such as the Average or Entropy.

Conversely, in the WM mappings the selected regions are different. When using Number of Folds and Thickness, the selected areas are located in the vicinity of those obtained in GM. However, our spherical maps, especially Average and Entropy, behave differently. There is still high t-values that correspond to the White Matter of the Parahippocampal gyrus, but large areas of negative t-values that are located in areas corresponding to the Caudate Nucleus, Globus Pallidus and Putamen. The areas corresponding to the Posterior Cingulate gyrus and adjacent Precuneus present also values related to cell loss, as suggested in [**Baron2001**].

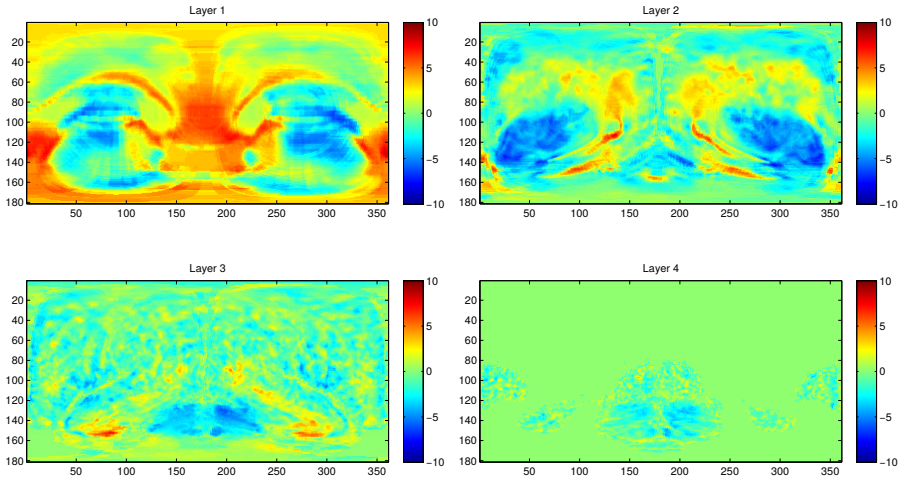
6.5.2.3 *Layered Extension*

The significance levels of the layered mappings has been assessed as well. However, due to space restrictions, we will only analyse the anatomical features of one of the mappings: a four-layered average mapping of the GM, that can be checked in Figure 6.9.

It is plain to see that most of the neurological changes in GM appear in layers 2 and 3, specifically in the Hippocampus, Parahippocampal lobe and Amygdala (layer 2) and the temporal lobe (layer 3), where the values of the average mapping (equivalent to the density of the tissue) are higher in normal control subjects than in AD affected patients. This reveals atrophy in these organs, as it has been previously reported in the bibliography [**Dubois2007**, **Pievani2013**]. In the case of WM, however, the changes are negative in the areas where the Rolandic Operculum, Heschl's gyri, Putamen and Globus Pallidus are found, and positive in some sections of the Hippocampus and the White Matter contained in the Parahippocampal lobe and the remaining parts of temporal lobe (layer 2 and 3). Nevertheless, the most significant differences are located in layer



a) GM



b) WM

Figure 6.9: t-maps that present the level of statistical relevance in the AD vs. NC paradigm, for a four-layered average mapping over a) GM and b) WM.

1, in the borders between ventricles and Thalamus, and specially in the Cuneus, Precuneus and Posterior Cingulate gyrus, which were reported in [**Baron2001**].

6.5.2.4 *VRLBP*

Finally, to end this statistical significance analysis, the t-maps of the more complex VRLBP mapping are presented in Figure 6.10.

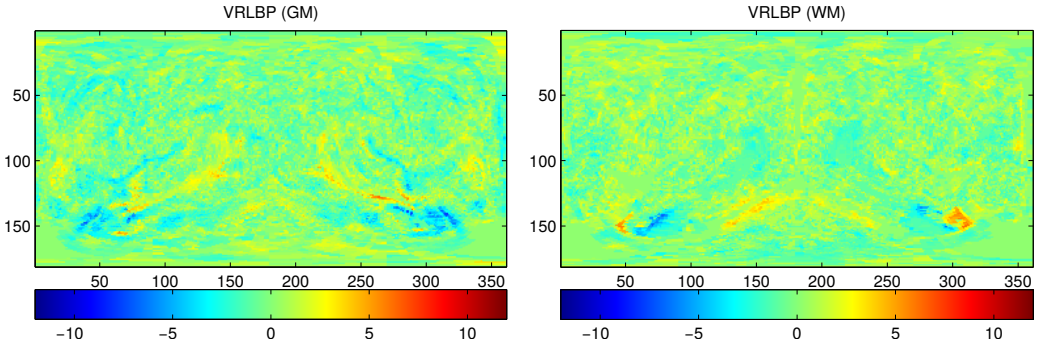


Figure 6.10: t-maps that present the level of statistical relevance in the AD vs. NC paradigm, for the VRLBP projections mapping over a) GM and b) WM.

These maps present low absolute t levels in most of the projection, however some small regions present high significance. These regions correspond to small areas in temporal lobe, Amygdala and Hippocampus in the GM, and even smaller regions in the WM corresponding to the limits between Hippocampus and Amygdala.

6.5.3 Classification Analysis

To obtain comparable performance metrics suitable to analyse the generalization capabilities of SBM, in this section a number of classification results are presented. A baseline is established in Section 6.5.3.1 and then the performance of our maps, included the layered extension and VRLBP, is presented in Section 6.5.3.2.

6.5.3.1 *Baseline - VAF*

In order to establish a baseline to assess the predictive ability of our maps, we will use the Voxels As Features (VAF) paradigm, described in [**Stoeckel04**]. This approach uses the whole 3D GM or WM segmented MR images and then uses all voxels of the 3D images as features in the SVM classification, yielding the

performance values shown in Table 6.1. The performance of the SBM maps will be compared to these.

Approach	Accuracy	Sensitivity	Specificity
VAF (GM)	0.768 ± 0.011	0.752 ± 0.016	0.785 ± 0.016
VAF (WM)	0.642 ± 0.009	0.668 ± 0.012	0.617 ± 0.013

Table 6.1: Performance values (Average $\pm$ Standard Deviation) for the Voxels as Features approach in both GM and WM tissues.

6.5.3.2 *Spherical Brain Mapping*

In this analysis, we have proceed as commented before, by computing the significance of each pixel using a t-test and then selecting a proportion of the most relevant, once they have been ranked according to their t-value. Later, these features are used to train and test a linear SVM classifier.

Approach	Perc.	Accuracy	Sensitivity	Specificity
Surface (GM)	0.100	0.638 ± 0.006	0.660 ± 0.030	0.616 ± 0.024
Surface (WM)	0.100	0.672 ± 0.007	0.692 ± 0.018	0.652 ± 0.018
Thickness (GM)	0.725	0.781 ± 0.007	0.811 ± 0.011	0.751 ± 0.017
Thickness (WM)	0.925	0.758 ± 0.009	0.773 ± 0.017	0.744 ± 0.011
Num.Fold (GM)	0.600	0.749 ± 0.013	0.782 ± 0.019	0.716 ± 0.013
Num.Fold (WM)	0.500	0.757 ± 0.005	0.745 ± 0.006	0.768 ± 0.009
Average (GM)	0.575	0.879 ± 0.005	0.897 ± 0.006	0.861 ± 0.006
Average (WM)	0.150	0.800 ± 0.011	0.802 ± 0.013	0.798 ± 0.009
Entropy (GM)	0.825	0.846 ± 0.008	0.842 ± 0.009	0.849 ± 0.011
Entropy (WM)	0.525	0.796 ± 0.006	0.811 ± 0.009	0.781 ± 0.009
Kurtosis (GM)	1.000	0.753 ± 0.007	0.801 ± 0.011	0.704 ± 0.015
Kurtosis (WM)	0.175	0.697 ± 0.008	0.702 ± 0.018	0.693 ± 0.009
VRLBP (GM)	0.200	0.903 ± 0.010	0.890 ± 0.012	0.916 ± 0.018
VRLBP (WM)	0.150	0.909 ± 0.014	0.899 ± 0.028	0.919 ± 0.018

Table 6.2: Performance values (Average $\pm$ Standard Deviation) for the different SBM approaches.

The results for each type of map are presented in Table 6.2, including the percentage of selected voxels (perc.) at which each value is obtained. Regarding

the Grey Matter, we can observe that the best type of mappings in the diagnosis task, in terms of average accuracy, are the Average (0.879 ± 0.005) and Entropy (0.846 ± 0.008). There results are followed by the measures of Thickness (0.781 ± 0.007), Kurtosis (0.753 ± 0.019) and the worse accuracy estimates are for Number of Folds (0.749 ± 0.013) and Surface (0.638 ± 0.006).

In the case of White Matter, and according to Table 6.2, the performance is again higher in Average (0.800 ± 0.011) and Entropy (0.796 ± 0.006). Thickness and Number of Folds present similar, but lower, perfomance values, respectively 0.758 ± 0.009 and0.757 ± 0.005, and being the Kurtosis (0.697 ± 0.008) and Surface (0.672 ± 0.007) maps the less powerful.

However, VRLBP outperform all these approaches by obtaining an accuracy of 0.903 ± 0.010 for GM and 0.909 ± 0.014 for WM, revealing itself as the best technique.

The evolution of the performance of the maps as the number of selected pixels varies is shown in Figure 6.11. In general, it is possible to see very small differences in the accuracy of the system, which makes its performance almost independent from the number of selected pixels. However, this is not the case of the Surface, and, more remarkable, the VRLBP. In the latter, the performance is the best for both tissues when the proportion of selected pixels is small, but degrades significantly as its number increases.

Regarding the four-layer extension to SBM, the performance values obtained by different mappings at different layers and thresholds (t-values of 2, 4, 8 and 10) is presented in Figure 6.12.

The first thing that we can observe is that for both GM and WM tissues, the better performance is achieved with the second layer. This is specially surprising in the WM case, as the highest t-values were located in layer 1.

6.5.4 Experimental Setup

In order to test our HMM-based path tracing algorithm, we propose the following experiments:

- Firstly, in Sec. 6.5.5 an evaluation of the algorithm over two synthetic 2D and 3D images.
- Secondly, in Sec. 6.5.6, the HMM paths created using the DARTEL template will be evaluated in the differential diagnosis (NC versus AD).
- Finally, in Sec. 6.5.7, the proposed texture feature maps computed along the DARTEL HMM paths will be evaluated in a differential diagnosis as well.

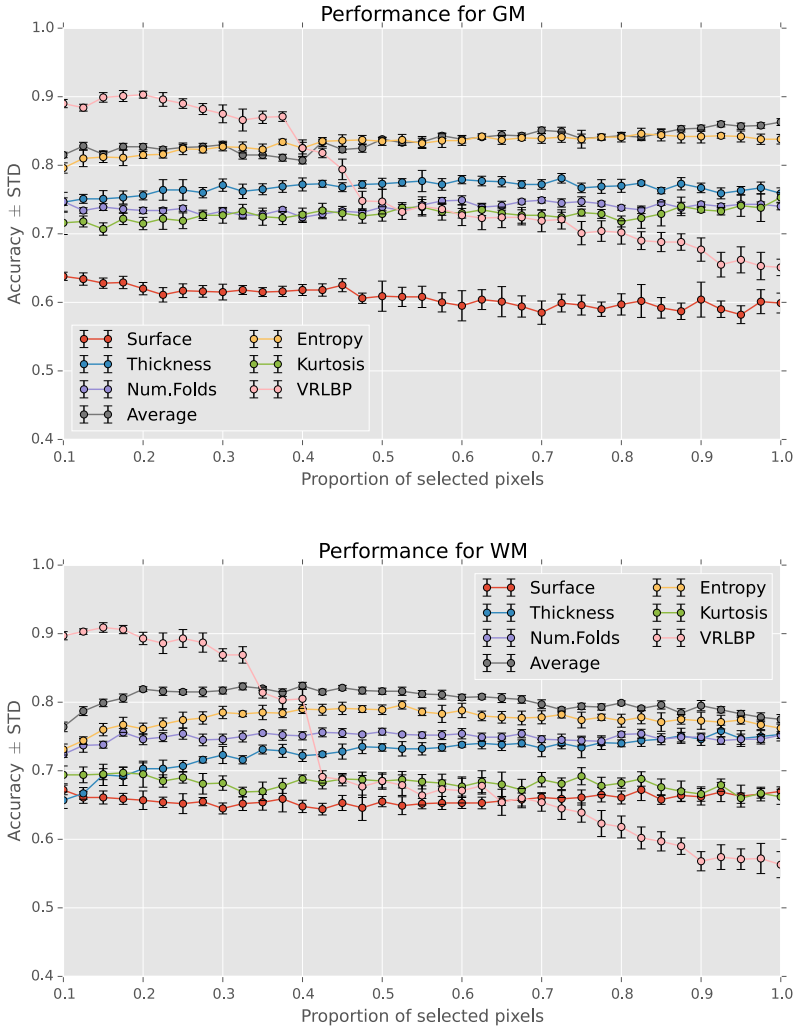
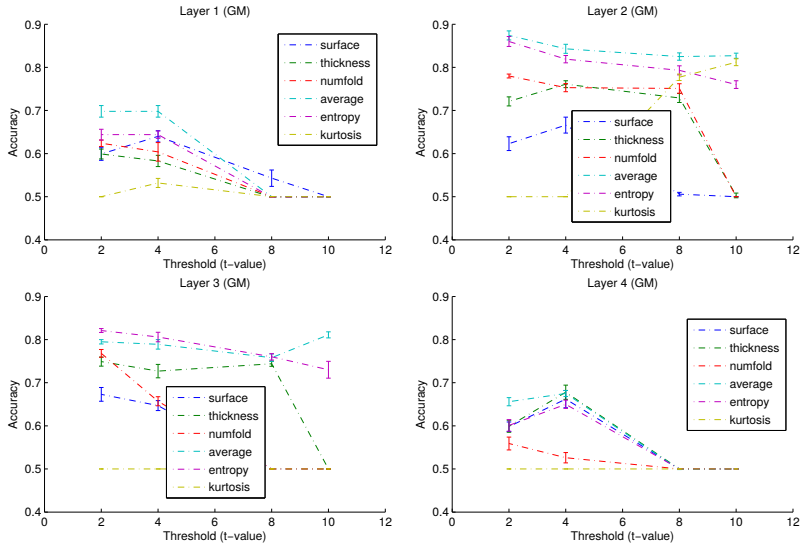
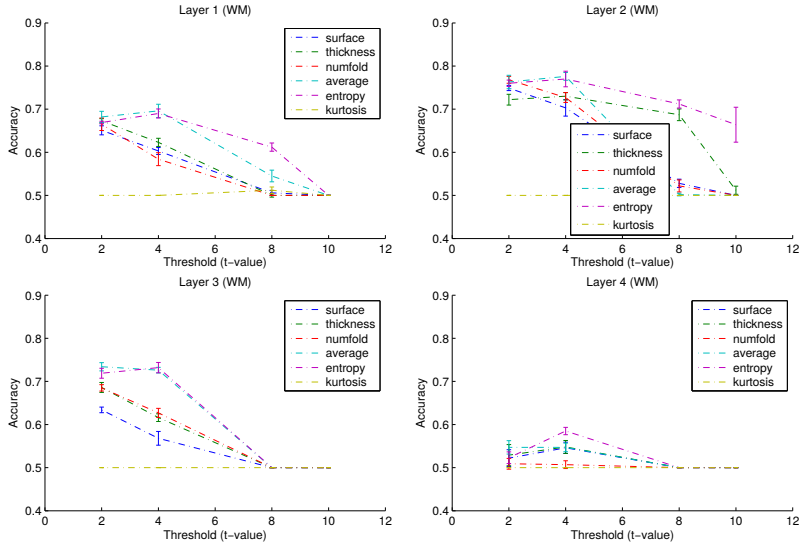


Figure 6.11: Performance for the different SBM approaches over the: a) Grey Matter and b) White Matter.



a) GM



b) WM

Figure 6.12: Performance for the different four-layered mappings over the: a) Grey Matter and b) White Matter at different levels of statistical significance.

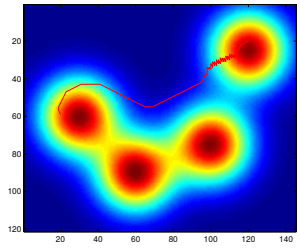


Figure 6.13: Path traced over a gaussian mixture distribution of 4 isotropic gaussian kernels.

For sections 6.5.6 and 6.5.7 a similar strategy is used to obtain performance results. Once a set of features has been extracted (intensity values in each path and statistical measures for Section 6.5.6 and texture features for Section 6.5.7), they are used to train and classify a Support Vector Machine (SVM) classifier with linear kernel, as implemented in `LIBSVM`[**Chang2001**], to classify the component scores. The classification was validated using stratified 10-fold cross-validation, as recommended in [**Kohavi1995a**].

6.5.5 2D and 3D demonstrations

A demonstration of the ability of our HMM path tracing algorithm can be found in Figures 6.13 and 6.14. In Fig. 6.13, the path tracing algorithm has been tested over a synthesized gaussian mixture probability density function using four isotropic gaussian kernels. The initial point was located at $\mathbf{x}_0 = (120, 20)$ and the attractor at $\mathbf{x}_N = (20, 60)$. The resulting path maximizes both the orientation of the path (towards $\mathbf{x}_N$) and the minimum change in the intensity values, which is specially visible in the last nodes of the path, where it approaches $\mathbf{x}_N$ surrounding the nearby kernel. In this case, the chosen L2-norm of the support ball has been $r = 3$.

The algorithm has been tested on a three-dimensional, helix-shaped point distribution as well (Fig. 6.14). The tracing algorithm needs per-voxel intensity (or probability) values, therefore we have estimated the probability distribution of the points as the number of points within each voxel over the total number of points. Using $\mathbf{x}_0$ as the point with minimum z coordinate in the data distribution and $\mathbf{x}_N$ the one with maximum z, the resulting path follows the data distribution consistently until it reaches the attractor.

Finally, we have tested the algorithm on a real world example, using a digital elevation model (DEM) of the Iberian Peninsula, generated by the LANDSAT

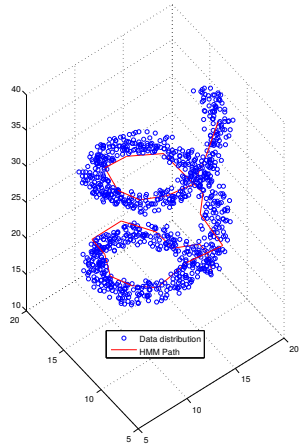


Figure 6.14: HMM path computed inside a density distribution defined by an helix.

SRTM30+ mission (see Fig. 6.15). We have tested a multiple path tracing by establishing sequentially $\mathbf{x}_0$ and $\mathbf{x}_N$ in ten cities. The resulting paths optimize both the distance and height variation, as well as resembling -in most cases- the roads that connect these cities in the real world. Given the dimensions of the image, in this case, the L2-norm of the support ball has been set to $r = 30$.

6.5.6 Intensity paths

In this section, we present the results of the first experiment involving paths in MRI. To do so, we define a set of canonical paths that are computed on the DARTEL template. These DARTEL paths model the anatomy of a normal subject to whom all other images have been registered. This means that we have fixed the location of the nodes to the structural information of the template, and by extension, to the general anatomy of all images in the database. Therefore, we can characterize the structural differences by the intensity distribution –in other words, the tissue density– of the voxels at the path nodes. Comparing the intensity distribution found in controls to the one found in AD affected subjects is thus the first logical step to measure how these paths can distinguish the different classes.

To test the algorithm we use the $180 \times 360 = 64800$ DARTEL paths computed in each spatial direction (φ, θ), with $\varphi \in [0, 360]$ and $\theta \in [-90, 90]$, to select the intensities in the voxels that are placed at the nodes. The amount of voxels selected ranges from 2 to several dozens. The set of selected intensities are used as features to train and test a SVM classifier. The accuracy reached by each

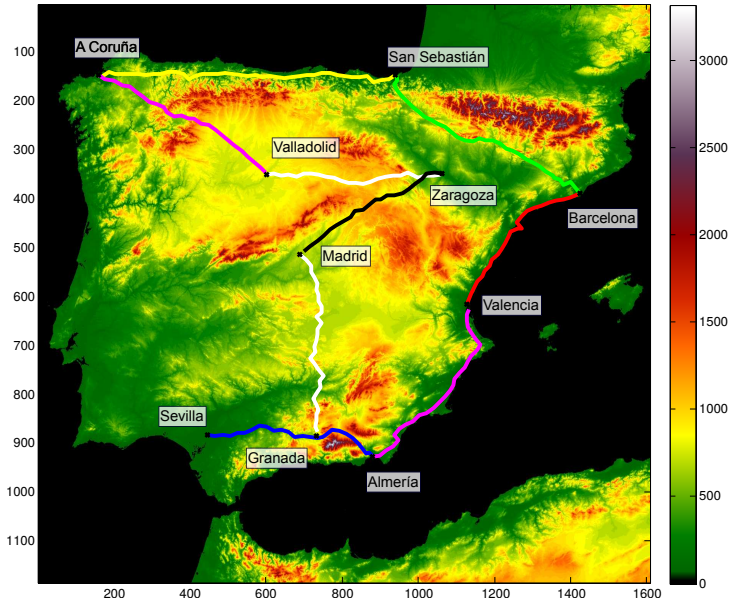


Figure 6.15: Simulation of the HMM-based path tracing over an Iberian Peninsula height map, interconecting different cities.

path (using the aforementioned cross-validation strategy) is presented as colour information in Figure 6.16. The higher accuracy obtained using only one path is 0.8028 ± 0.0873, and corresponds to the light green paths that cross the temporal lobe.

It is interesting to question if the performance of this differential diagnosis could be improved using the information contained in more than one path at a time. To this end, we first take the higher accuracy (accuracy $\geqslant 0.7$) paths according to the aforementioned performance and select all voxels located in the nodes of these paths. Additionally, we use a t-test over the set of voxels selected by these paths, to further reduce the set to those voxels that have significant ($p < 0.05$) t-values ($|t| > 1.96$). The performance values for the experiment involving all voxels in the paths (first row) and the one that uses only those significant voxels (second row) are presented in Table 6.3.

Finally, we mimic the procedure followed in the SBM article[**Martinez-Murcia2015**]. That is, we first compute the average, variance, entropy and kurtosis maps of each brain, but instead of using rectilinear paths, we use the DARTEL paths. Afterwards, all the features contained in these maps are used as an input to the SVM classifier. The performance results are shown in Table 6.4.

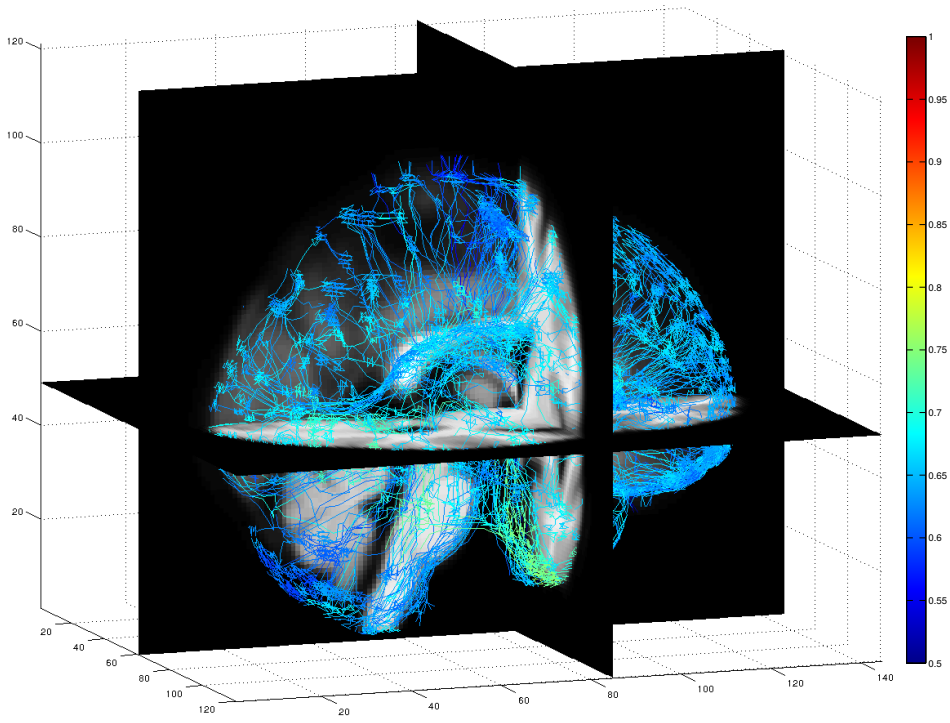


Figure 6.16: DARTEL paths computed in each direction (φ, θ). Each path's colour represent the accuracy in a differential diagnosis. Only one in every five paths are shown for clarity purposes.

Side	Accuracy	Sensitivity	Specificity
Both	0.806 ± 0.069	0.733 ± 0.073	0.878 ± 0.097
Left	0.769 ± 0.035	0.717 ± 0.061	0.822 ± 0.057
Right	0.792 ± 0.080	0.706 ± 0.120	0.878 ± 0.101
Both	0.828 ± 0.054	0.794 ± 0.095	0.861 ± 0.039
Left	0.733 ± 0.037	0.694 ± 0.099	0.772 ± 0.124
Right	0.781 ± 0.085	0.711 ± 0.122	0.850 ± 0.083

Table 6.3: Performance values ($\pm$SD) for the selected paths as features, and using t-test to select the voxels.

Feature	Accuracy	Sensitivity	Specificity
Average	0.594 ± 0.062	0.661 ± 0.121	0.528 ± 0.106
Variance	0.750 ± 0.064	0.633 ± 0.131	0.867 ± 0.102
Entropy	0.603 ± 0.069	0.661 ± 0.071	0.544 ± 0.125
Kurtosis	0.756 ± 0.105	0.733 ± 0.165	0.778 ± 0.150

Table 6.4: Performance values (±SD) for each of the measures used in the SBM article.

Feature	Accuracy	Sensitivity	Specificity
Contrast	0.733 ± 0.060	0.689 ± 0.126	0.778 ± 0.105
Correlation	0.672 ± 0.068	0.672 ± 0.112	0.672 ± 0.100
Dissimilarity	0.711 ± 0.085	0.678 ± 0.110	0.744 ± 0.102
Energy	0.689 ± 0.061	0.700 ± 0.115	0.678 ± 0.073
Entropy	0.675 ± 0.101	0.672 ± 0.115	0.678 ± 0.159
Homogeneity	0.697 ± 0.058	0.700 ± 0.115	0.694 ± 0.106
Difference Variance	0.736 ± 0.070	0.683 ± 0.098	0.789 ± 0.090
Difference Entropy	0.725 ± 0.122	0.683 ± 0.176	0.767 ± 0.114
IDN	0.719 ± 0.065	0.683 ± 0.108	0.756 ± 0.105
IDMN	0.717 ± 0.076	0.678 ± 0.125	0.756 ± 0.084

Table 6.5: Performance values (±SD) for each of the 10 texture features.

6.5.7 Texture features

The second experiment is intended to extract texture features from the DARTEL paths. With this approach, we obtain one single value per texture feature and path in the subjects, values that intrinsically contain information from their location in the path, in contrast to standard SBM measures. In the end, each subject will be characterized by a 2D, 361×181 array of scalars, one for each texture feature applied to the paths. Performance values for the nine texture features maps from Section 6.4.1 are presented in Table 6.5.

The higher accuracy obtained by the texture maps is 0.736 ± 0.070, corresponding to Difference Variance. The performance values of the different texture features, all obtaining accuracies higher than 65% (most of them above 70%) reveal the discrimination abilities of these textures, although these values are not as good as those obtained using the voxel intensities or the SBM features.

6.6 Discussion

6.6.1 Spherical Brain Mapping

The structural changes in MR images during the progression of the Alzheimer's Disease are widely documented in the bibliography [**Misra2009**, **Baron2001**, **Pievani2013**, **Stoeckel04**, **han2006reliability**, **Fischl2004**]. According to our current knowledge, the neurodegeneration and posterior atrophy occurs mainly in the GM tissue, although significant changes are present also in WM.

The mappings defined throughout Sections **??**, 6.2.1 and 6.3 account for different properties of the tissues crossed by $\mathbf{v}_{\theta,\varphi}$. As it can be seen in Figure 6.17, our mappings show in general a higher performance when using the GM tissue, which is consistent with the literature. There are some exceptions, however, being the clearest the VRLBP, and, to a lesser extent, the number of folds and surface. The different mappings and their utility will be described in the following paragraphs.

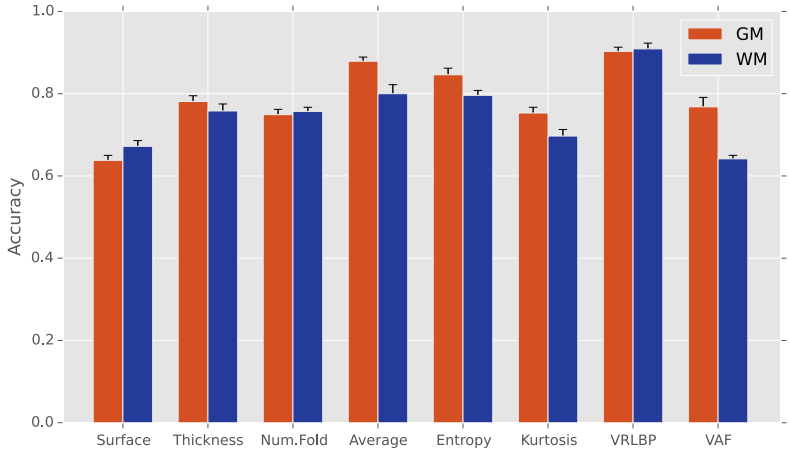


Figure 6.17: Performance at the operation point for the different mappings over the Grey Matter and White Matter, compared with the performance of VAF.

The first three approaches, Surface, Thickness and Number of Folds are easily interpreted, as they intend to represent the surface of the tissue by mapping the distance between the centre of the image and the last voxel, the thickness of the tissue, and a measure of the complexity of the different sulci and gyri.

Surface and Thickness are highly related to other measures provided by widely-used software. However, as they are related to our more general SBM description,

their performance is poor, specially in the case of the Surface mapping. As it can be seen in Fig. 6.3, and later in the t-maps at Fig. 6.8, the detail of the surface map lacks higher detail, specially due to the superposing gyri and sulci. These superposition occur to a lesser extent in WM tissue, and this is probably why this technique obtains higher performance in WM than in GM.

As for the case of Thickness, although similar, it gathers much more information than the surface, without achieving, however, the level of detail of the cortical thickness measures provided by Freesurfer [**Fischl2004**] or other software. Nevertheless, cortical thickness it is a descriptive, widely accepted as a measure of neurodegeneration in Alzheimer's Disease in the literature [**han2006reliability**, **Fischl2004**], and its measures might be relevant for a subsequent analysis.

Number of Folds, however, is intended to model the complexity of the cerebral cortex, and therefore, it is of far more use in the case of GM than in the WM. This can be easily checked when looking at the maps obtained for both GM and WM in Figure 6.3.

The last three measures described in Section **??** are statistical values that describe the variability of the sampling set $V_{\theta,\varphi}$. It would be reasonable to expect the better performance to be linked to the mapping that better models the tissue atrophy.

This is the case of the average of these intensities, which can be interpreted as the total amount of tissue, being therefore a good measure of the level of brain atrophy in each direction (θ, φ). The average maps show the best performance of all the measures proposed in Section **??**, and is higher in GM than in WM. This is consistent with the literature, as atrophy mainly occurs in GM tissues.

Entropy is a more complex statistical concept that comes from information theory, but is usually related to the amount of information, or in other words, the "randomness" of a source. In our particular case it could be interpreted as a measure of texture, that is, the grey-level variability in the direction of $\mathbf{v}_{\theta,\varphi}$. These maps perform very similar to the average ones in both GM and WM, suggesting that the entropy accounts for the tissue density as well.

The last mapping defined, Kurtosis, is a fourth-order statistic, often interpreted as the peakedness (width of peak) of a probability distribution. In our context, it is related to the sharpness of the changes in the direction of $\mathbf{v}_{\theta,\varphi}$, and thus is related to the number of folds. As in the case of the latter, the Kurtosis performs poorly in both types of tissues, probably because they are measures that are not as directly related to atrophy as other measures such as average, entropy or thickness.

The last of the single measures proposed in this work is the Volumetric Radial LBP defined in Section 6.3. It is a measure of the texture not only in the direction

of $\mathbf{v}_{\theta,\varphi}$, but also in the neighbourhood of the mapping vector. Therefore, it is not strange that it obtains the best performance of the whole work, yielding accuracy results above 0.9 for both GM and WM tissues.

This could seem counter-intuitive, as the t-maps for this technique, presented in Fig. 6.10, show small regions of high significance, when compared to the measures in Sec. **??**. Yet, despite its size, it performs fairly well with a relatively small amount of data. It is probably due to the nature of VRLBP, and the areas highlighted in Fig. 6.10 probably correspond to the texture changes associated to the loss of tissue in the Hippocampus.

As for the layered extension, which might seem a powerful method to add detail to the mappings, obtains however similar performance to the methodology above. It seems that the amount of information that can be obtained by each measure does not depend on the number of layers, and accordingly, its benefits are only related to visualization. In this case, best values are obtained in layer 2, which is consistent to the presence of some organs, specially the Hippocampus.

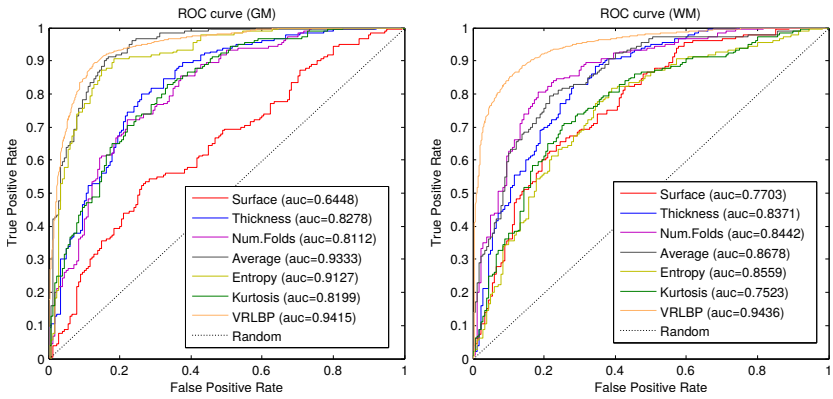


Figure 6.18: ROC curves of the different mappings for the GM and WM tissues.

Finally, in order to have another look at the performance of our mappings, the ROC curves of each type are presented in Figure 6.18. There we can see how the VRLBP approach outperforms all the other measures, specially in the case of WM tissue. In GM, Average and Entropy present values really close to VRLBP, as expected. Conversely, the poorest performance is achieved by the Kurtosis and Surface mappings, however the Surface performs better in WM than in GM. These results confirm the performance values presented in Table 6.2 and Figure 6.11, making our proposed mapping framework a reasonable choice for obtaining both a visual interpretation of otherwise hidden features and a significant dimensionality reduction.

It is important to note that our Spherical Brain Mapping defines a whole framework that can be easily extended with different sampling strategies. This

is the case of the layered extension and the helical sampling in VRLBP, but they are only two examples of what can be done. Since our simplest approach implies a computation of a value from a vector of intensities, measures used to describe time-course data could be added to complete and highlight different properties of the tissues. In this context, high-order statistics [**Zhou2008**], as well as spectral measures [**Locatelli1998**] have been successfully applied to analyse electroencephalogram (EEG) signals, and could be therefore applied here to bring different structural properties of the images into focus. Additionally, our mapping method is potentially applicable to other imaging modalities, such as PET and SPECT, where the structural information is sometimes lost [**IAIllan2010**, **Ram'irez2009**]. Our technique does not need the use of complex co-registering of MRI and functional imaging to locate cerebral structures, as it rely only in their angle and depth. Moreover, in the case of Diffusion Tensor Imaging (DTI), which has proven itself as a good tool for the diagnosis of Alzheimer's Disease [**Grana2011**, **Medina2008**], SBM could be modified to replace $\mathbf{v}_{\theta,\varphi}$ with each tract, and subsequently project a given feature, resulting in a summary of the tract's behaviour in a single two-dimensional image.

6.6.2 Paths via HMM

In this work we propose a new path tracing algorithm based on Hidden Markov Models used to trace similar intensity paths inside the brain. The paths are meant to be used as a feature extraction tool in the SBM framework either by selecting voxels or computing features. We have performed several experiments to evaluate these approaches in a differential diagnosis of AD using MRI brain images.

Our paths are defined so that they construct a minimum intensity variation path starting at the AC and oriented in a general direction set by the spherical coordinate pair (φ, θ). As commented before, the AC is the obvious starting point, given its privileged position in the middle of the left and right hemispheres. A different starting point will reveal suboptimal, stopping at disconnected regions such as the ventricles, and yielding incomplete paths.

The paths adapt to the intensity changes in a certain direction in the brain, modelling grey level connectivity in all spherical directions. Since grey level is directly related to tissue density, we can assume that the outcome follows smooth, same-density paths that start in white matter and progressively transition to grey matter in a specific direction. Therefore, they are not functional connectivity maps like Diffusion Tensor Imaging (DTI), which have been used as well in the diagnosis of AD[**Grana2011**, **Medina2008**]. While DTI fibers are

the result of a tensor processing over diffusion images that quantify the water molecule motion -in both direction and average magnitude- at the voxel level, our HMM paths only characterize grey level connectivity in static MRI images, and are meant to be used for feature selection.

Our first experiment uses HMM paths computed on the DARTEL template to describe how the intensity of the set of voxels corresponding to a certain path can be used as discriminant features in a SVM classifier. The differences in the distribution of intensities between controls and AD affected subjects are used to identify structural changes in AD. Fig. 6.19 depicts the paths that achieved best performance (accuracy higher than 0.75) in this differential diagnosis, superimposed to some structures rendered from the Automated Anatomical Labeling (AAL) brain atlas[**Tzourio-Mazoyer2002**].

The paths that obtained higher accuracy are those that cross structures such as the Hippocampus, Amygdala, Thalamus, Fusiform and Inferior Temporal Gyrus. Particularly, grey matter loss in the Hippocampus has been described in the NINCDS-ADRDA criteria for AD diagnosis[**Dubois2007**] and is widely accepted[**chan2001patterns**, **Baron2001**, **Jong2008**]. Furthermore, the evidence suggest that atrophy affects the surrounding structures (Amygdala, Parahippocampal and Fusiform Gyrus) as well[**chan2001patterns**, **Baron2001**]. Some studies have found significant atrophy in the Thalamus and Putamen in early AD[**Jong2008**] as well. Generally, in advanced AD, most of the neocortex and grey matter suffer from atrophy[**chan2001patterns**, **Baron2001**, **Jong2008**], which explains why most of the paths that involve the neocortex in Fig. 6.16 obtain accuracy rates around 0.7.

A number of feature maps have been computed as well. These are the result of applying some of the SBM measures to the voxels selected by the HMM paths. Variance and kurtosis have been proved as the most discriminative maps (with accuracy higher than 0.7). This is coherent with the definition of the paths, where the intensity transitions are minimal. Therefore, average would be the less discriminative in this case, being higher order statistics such as variance or more representative of the tissue density distribution of each class.

Regarding texture analysis, we have again discriminative features (with accuracy that surpass the 70%) yet not very powerful. This situation might be due to the definition of the paths as minimum intensity variation paths, being the textural changes along the path minimal.

However the real utility of these texture features could be in its application to longitudinal studies, since textsure can be related to evolution of the disease[**sikio2015mr**]. It is very convenient to use a scalar to characterize a measure (in our case, texture features) in each direction. The texture obtained in each session can be used to construct a function of neurodegeneration that al-

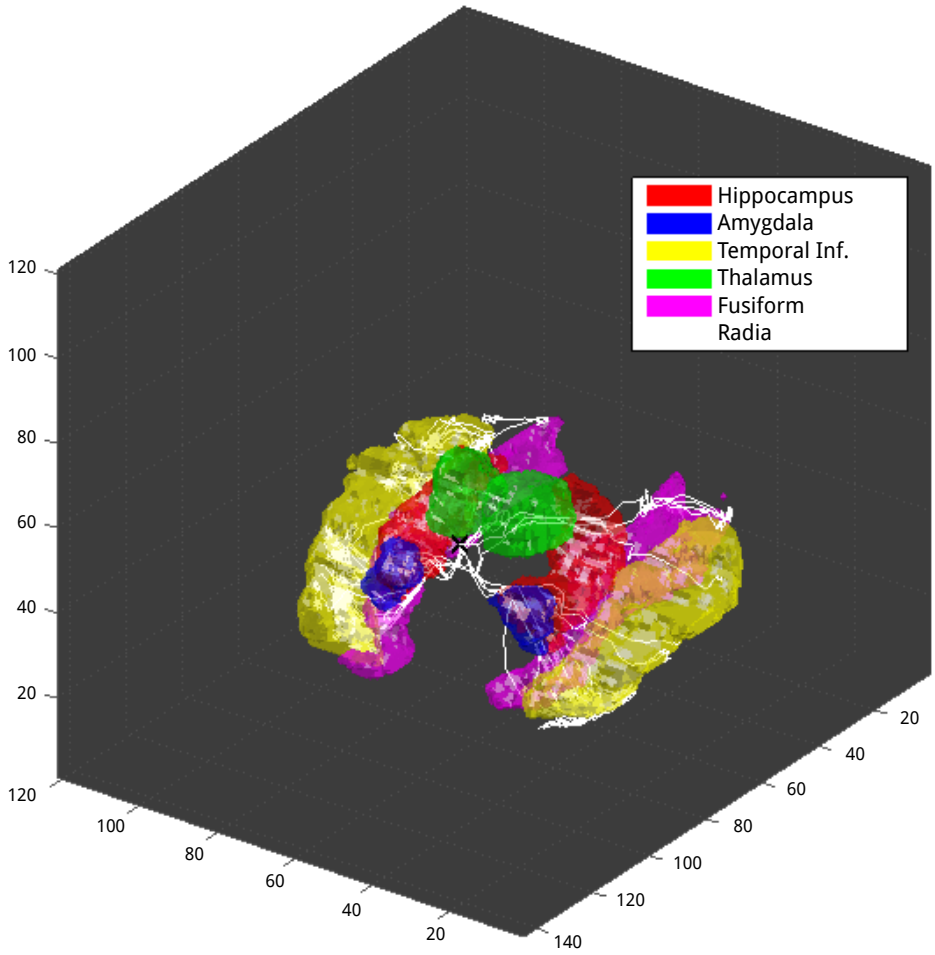


Figure 6.19: Paths that obtain more than 75% accuracy, and a three-dimensional representation of the structures crossed by them.

Feature	Accuracy	Sensitivity	Specificity
Paths	0.806 ± 0.069	0.733 ± 0.073	0.878 ± 0.097
Selected Paths	0.828 ± 0.054	0.794 ± 0.095	0.861 ± 0.039
Variance	0.750 ± 0.064	0.633 ± 0.131	0.867 ± 0.102
Kurtosis	0.756 ± 0.105	0.733 ± 0.165	0.778 ± 0.150
Texture (Difference Variance)	0.736 ± 0.070	0.683 ± 0.098	0.789 ± 0.090
VAF	0.768 ± 0.011	0.752 ± 0.016	0.785 ± 0.016
SBM-average (GM)	0.879 ± 0.005	0.897 ± 0.006	0.861 ± 0.006
SBM-average (WM)	0.800 ± 0.011	0.802 ± 0.013	0.798 ± 0.009
SBM-VRLBP (GM)	0.903 ± 0.010	0.890 ± 0.012	0.916 ± 0.018
SBM-VRLBP (WM)	0.909 ± 0.014	0.899 ± 0.028	0.919 ± 0.018
LVQ-SVM (GM)	0.869 ± 0.101	0.822 ± 0.120	0.890 ± 0.102
SCA (GM)	$0.880 \pm 0.0^*$	$0.926 \pm 0.0^*$	$0.845 \pm 0.0^*$
SCA (WM)	$0.808 \pm 0.0^*$	$0.817 \pm 0.0^*$	$0.800 \pm 0.0^*$

* SCA used leave-one-out cross-validation. SD is 0.

Table 6.6: Comparison between our algorithm performance values (best values for selected voxels in all paths and texture features) (±SD) and other methods in the bibliography

lows the exploration of the different stages of the disease as the changes in the brain texture along the time within a single patient.

Table 6.6 presents some of the best results of our methodology involving HMM paths in this order: the performance of a single path and using the selected paths as features (Section 6.5.6), the performance of using projected maps (in this case, variance and kurtosis) like in the SBM paper (Section 6.5.6) and the results of computing texture maps using radial GLCM and Haralick Texture Features (Section 6.5.7). It is compared with the methods using in the SBM paper[**Martinez-Murcia2015**], the SBM-VRLBP[**Martinez-MurciaVRLBP**], the Voxels As Features (VAF)[**Stoeckel04**] algorithm and different approaches used in the ADNI database and involving SVM classifiers such as the LVQ-SVM[**Ortiz2013**] or Spatial Component Analysis (SCA)[**Illan2014**].

VAF is often used as a baseline when comparing different methodology, as it has been described as a good estimator of the accuracy obtained by means of visual analysis [**Stoeckel04**]. As we commented before, the raw voxel intensities selected by our DARTEL paths achieve higher accuracy than statistical or texture features, and it is the only strategy that outperforms VAF. Texture and statistical features obtain poorer, although still good, performance (around 75% accuracy). When compared to other methods, the difference is greater, although inside the

range of 1 SD. Most of the SBM features proposed in [**Martinez-Murcia2015**] perform better than our DARTEL paths, and the case of [**Martinez-MurciaVRLBP**] even surpass the barrier of 90% accuracy. However, there is a significant difference with these approaches, and it is that these measures used segmented GM and WM images, instead of using the whole MRI. Segmentation, thus, enhances the detection and extraction of features from the images, whereas the tracing of paths over the whole images is a more complex operation. When compared to LVQ-SVM or SCA, the difference in performance is even smaller and still inside the range of 1 SD, which gives us an idea of the ability of our methodology to detect patterns with a significant feature reduction.

Finally, one might argue if a different approach to the path tracing, such as tracing the set of paths in each subject individually might be of use. This strategy would still characterize the individual brain structure; however the way this structure is defined would be different: the spatial location of the nodes and topology of the paths instead of the intensity distribution. Given the time our algorithm takes to model one single MRI (around 2 hours) it can be extraordinarily computationally expensive, although faster than other methodology like DTI fiber tracing or Freesurfer surface extraction. Consequently, it would be an interesting option to explore in future works.

Part III

INCREASING THE SAMPLE SIZE

7 SIGNIFICANCE WEIGHTED PRINCIPAL COMPONENT ANALYSIS

Multicentre studies with structural (sMRI) and functional Magnetic Resonance Imaging (fMRI) are increasingly common, allowing for recruitment of larger samples in shorter periods of time. However, the use of images acquired at different sites still poses a major challenge. In addition to logistical difficulties, such as regulatory approvals and data protection, a number of technical and methodological issues can potentially affect the resulting maps, introducing undesired intensity and geometric variance. This issue has been addressed in other neurological conditions, such as Alzheimer's Disease (Jovicich, et al., 2006; Stonnington, et al., 2008), where group differences are well known, and demonstrating that the impact of a correction for site on the resulting neurobiological differences is relatively small. However, these effects have a stronger impact in psychiatric conditions where the atypical radiological signs on MRI are often subtle and require large samples of patients to observe on-average differences relative to control samples. Recent meta-analyses point to differences being inconsistently reported in schizophrenia (Friedman and Glover, 2006; Turner, et al., 2013), psychosis (Clementz, et al., 2016; Wang, et al., 2015), and ASD (using the multi-centre ABIDE database) [**haar2014anatomical**]

These inconsistencies can arise from a variety of variance sources, ranging from the multi-level (phenotypic, neurobiological, and etiological) heterogeneities of the conditions to technical issues that include differences in scanner make, model, manufacturer, static field strength, field inhomogeneities, slew rates and image reconstruction (Van Horn and Toga, 2009), as well as acquisition problems such as within-acquisition participant head motion. Field inhomogeneities are a source of misinterpretation of the data even when the same MRI system manufacturer and model are used (Van Horn and Toga, 2009). Furthermore, results in (Pearlson, 2009) demonstrate that a single scanner can change with time, which makes some widely used strategies, for example collecting controls first and patients later, a flawed approach. Recent neuroimaging research on ASD (Haar, et al., 2014) has shown that, while analyses performed on a particular database (acquired on a single platform) could yield coherent regions, the atypical structures are often inconsistent across the wider literature using different databases. Therefore, new methodologies focused on reducing multi-

site variance may be potentially helpful in increasing the power to identify the characteristic neurobiological signature of autism, should there be one.

[Martinez-Murcia2016a]

7.1 Significance Weighted Principal Component Analysis

The Significance Weighted Principal Component Analysis (SWPCA) is an algorithm to reduce, in this case, undesired intensity variance introduced by multi-site image acquisition. SWPCA takes any dataset of pre-processed images, spatially normalized, and decomposes them into their variance components to then provide a corrected dataset where these undesired variance components have been reduced. To do so, PCA was applied to each modality in turn to obtain the component scores and component loadings. Since PCA is a data-driven approach, it was only used to decompose the source images, and after this procedure, a one-way ANOVA estimated the relation between each variance component and a given categorical variable, in our case, the acquisition site. The between-site variability in the variance component was then identified by its corresponding p-value. Finally, these p-values were transformed into a weighting matrix $\mathbf{\Lambda}$ that weighted the influence of each variance component in a final PCA reconstruction of the corrected maps. The procedure is summarized in Figure 7.1.

7.1.1 Principal Component Analysis

The first step in the SWPCA algorithm was to perform a PCA decomposition of the dataset into a set of orthogonal components that model the variance present in the images.

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations $\mathbf{X}$ of possibly correlated variables, where $\mathbf{X}$ is a $K \times N$ matrix, with K participants (in this case, with one image per participant) and N the number of voxels, into a set of N linearly uncorrelated variables called Principal Components (PC, also known as component loadings or the mixing matrix) $\mathbf{W}$ of size $N \times N$ whose linear combination using a vector of component scores s_K can perfectly recompose each image. The set of these component scores $\mathbf{S}$ (size $K \times N$) was estimated as:

$$\mathbf{S} = \mathbf{X}\mathbf{W}^T \tag{7.1}$$

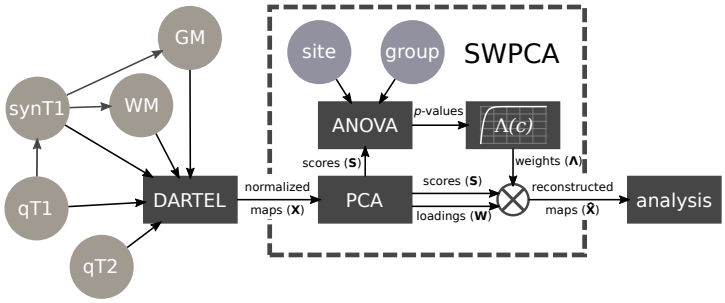


Figure 7.1: Summary of the SWPCA algorithm, along with its context in the pipeline used in this article. Circles represent the input data, both images (green shading) and class (group and acquisition site, purple shading). Rectangles represent the different procedures applied, comprising the DARTEL normalization and registration, the different steps contained in SWPCA, ANOVA and obtaining the weighting function $\Lambda(c)$- and the subsequent analysis.

This transformation computes a sequence of PCs, maximally explaining the variability of the data while maintaining orthogonality between components. PCA was computed using Singular Value Decomposition (SVD):

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{*} \tag{7.2}$$

where U is an $K \times K$ orthogonal matrix, $\mathbf{\Sigma}$ is a $K \times N$ diagonal matrix with non-negative real numbers on the diagonal, and the $N \times N$ unitary matrix $\mathbf{V}^{*}$ denotes the conjugate transpose of the $N \times N$ unitary matrix $\mathbf{V}$. With this decomposition both the component scores and estimates of the set of components loadings $\mathbf{W}$ were obtained. In this work the truncated form of SVD was used such that only the first C components were considered, where most of the variability of the data was concentrated:

$$\mathbf{S}_C = \mathbf{U}_C\mathbf{\Sigma}_C = \mathbf{X}\mathbf{W}_C \tag{7.3}$$

where $\mathbf{S}_C$ is the set of component scores using the first C components (size $K \times C$). To achieve reasonable performance with minimal information loss, it was assumed that the number of components was the same as the number of images, $C = K$. Thus, a partial reconstruction of the original signal could be undertaken:

$$\hat{\mathbf{X}} = \mathbf{S}_C\mathbf{A}_C \tag{7.4}$$

where $\mathbf{A}_C$ is the pseudoinverse of the truncated matrix of component loadings $\mathbf{W}_C$, and $\hat{\mathbf{X}}$ is the reconstructed set of images.

7.1.2 One-Way Analysis of Variance

The estimated PCs effectively model the variability of the image dataset. The next step was to assess each PC as a source of inter-site variance with one-way Analysis Of Variance (ANOVA). ANOVA estimates the F-statistic, defined as the ratio between the estimated variance within groups and the variance between groups:

$$F = \frac{MS_{within}}{MS_{between}} = \frac{SS_{within}/(G-1)}{SS_{between}/(K-G)} = \frac{\sum_i n_i \left(\bar{Y}_i - \bar{Y}\right)^2/(G-1)}{\sum_{ij} \left(Y_{ij} - \bar{Y}_i\right)^2/(K-G)} \tag{7.5}$$

Where MS_{within} and $MS_{between}$ are the mean squares within- and between-groups respectively, G is the number of separate groups (in our case, two), $\bar{Y}$ is the sample mean of a certain feature (in our case, the sample mean of all K values of a given component score), $\bar{Y}_i$ is the sample mean of the features belonging to group $i = 1 \dots G$, Y_{ij} is the j_{th} observation of a feature belonging to group i and n_i is the number of participants in the i_{th} group. The F-distribution allows an easy computation of p-values, given the number of groups and degrees of freedom. The F-statistic and p-values were computed independently for each component score and acquisition site, and then used in the SWPCA algorithm.

7.1.3 Weighting Function

To obtain a set of corrected maps, a new signal matrix of all maps of the same modality, $\widehat{\mathbf{X}}$, was estimated with the influence of the PCs with variance related to acquisition site, assessed via the p-values, reduced. To do so, equation 7.4 was modified to include a square matrix $\mathbf{\Lambda}$ (dimension $C \times C$) whose diagonal contains a weight λ_c for each component that depends on its p-value; that is,

$$\widehat{\mathbf{X}} = \mathbf{S\Lambda A} \tag{7.6}$$

The computation of each λ_c, for each component, was performed using the Laplace distribution, modified so that the weights were on the interval [0, 1]:

$$\Lambda_c(p_c, p_{th}) = 1 - e^{\frac{-p_c}{p_{th}}} \quad \forall p_c \in [0, 1] \tag{7.7}$$

where p_c is the statistical significance of the c_{th} component with respect to the acquisition site and p_{th} is the statistical threshold for significance; that is, p_{th}=0.05. A plot of the univariate weighting function $\Lambda_c(p_c, p_{th})$ can be found in Figure 7.2. This weighting ensured that most of the components of variance that are not related to the acquisition site are kept unchanged, while at the same

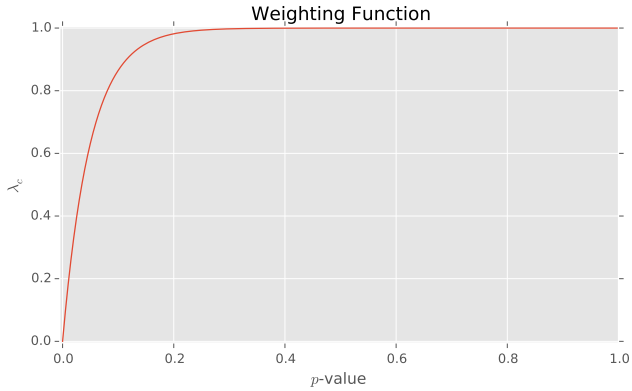


Figure 7.2: Weighting function $\Lambda_c(p_c, p_{th})$ used in SWPCA.

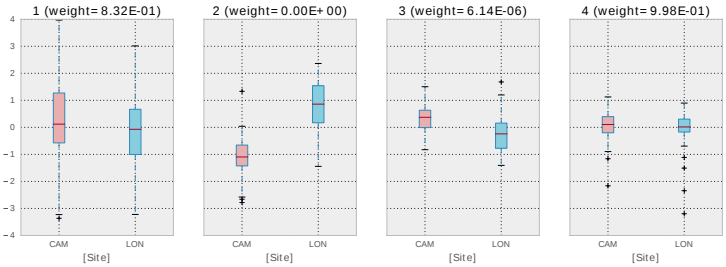


Figure 7.3: Box-plot of the distribution of the component scores at each site of the AIMS-MRI dataset (see Sections 7.2 and A.1.2) in the four first components. We assume that bigger differences between distributions imply a bigger influence of the acquisition site on the portion of variance modelled by that component and therefore, to parse out those differences, the resulting weight will be smaller.

time it strongly reduces the influence of components with p-values less than the threshold.

This procedure is illustrated in Figure 7.3, where a boxplot of the distribution of the first four principal component scores is shown. Since we have assumed that substantial differences imply a bigger influence of the acquisition site on the portion of variance modelled by that component, the resulting weight is reduced, and the contribution of that component to the reconstructed signal will be smaller. After computing all weights, most of the sources that are related to the acquisition site (for example, the second and third components) have been parsed out while keeping all other sources of variance.

7.2 Results for AIMS-MRI Dataset

To validate the effects of the SWPCA algorithm on the inter-site variance, experiments were undertaken to assess the reduction of the undesired site variance in the original datasets, and its impact on the between-group signal. Two kind of analysis were performed: a characterization of voxel-wise differences, and a classification analysis.

Voxel-wise differences between groups were characterized using Voxel Based Morphometry (VBM) [**Ashburner2000**], comprising preprocessing (registration, smoothing) and mass-univariate t-test on the smoothed maps from each modality. SWPCA is included (when needed) in this pipeline as a plug in, after the smoothing and before the computation of the test. Permutation testing assessed the significance of the relationship between the tested and target variables. A max-type procedure was used to obtain family-wise, whole-brain corrected p-values (Freedman and Lane, 1983). Additionally, a Component Based Morphometry (CBM), based on Source Based Morphometry (SBM) [**xu2009source**] was used. This procedure provided Z-maps for visual inspection comparable to those obtained in VBM, by selecting component loadings $\mathbf{W}$, scaling them to unit standard deviation and weighting their contribution to the final map with their statistical significance, computed using the same permutation inference as in VBM.

A classification analysis was undertaken using a common classification pipeline (Khedher, et al., 2015; López, et al., 2009) consisting of preprocessing, feature extraction and classification. SWPCA is used as a plug-in here as well, after the preprocessing and before the feature extraction step. We used PCA on the images for feature reduction and a Support Vector Classifier (SVC) with linear kernel, as implemented in LIBSVM [**Chang2001**], to classify the component scores in both corrected and uncorrected datasets (i.e. with and without SWPCA).

The classification was validated using stratified 10-fold cross-validation [**Kohavi1995a**]. In brief, 9 subsets of the dataset were used for extraction of the PCs and training of the classifier with the remaining subset used for testing. This procedure was repeated for each subset, repeated 10 times to avoid possible bias and random effects of the partitions. The average and standard deviation of the accuracy (acc), sensitivity (sens) and specificity (spec) values for each repetition were recorded.

For each modality independently, the following experiments were performed:

- **Experiment 1**: To demonstrate the ability of the SWPCA algorithm to reduce undesired effects due to acquisition site, the PCA + SVC pipeline was applied to the datasets labelled by acquisition site. Classification accuracy was compared to datasets with and without SWPCA. VBM was then ap-

plied to identify the spatial location of the between-site differences. This was undertaken on the whole database (ALL), and subgroups containing only ASD or ASD participants.

- **Experiment 2**: The discrimination ability of each modality, acquired at different sites was assessed by classification performance of individuals from London (LON) and Cambridge (CAM) was separately assessed, using group (ASD and CTL) as the labels.

- **Experiment 3:** To assess the impact of SWPCA on the datasets when characterizing the differences between ASD and CTL groups, the classification pipeline comprising PCA + SVC, as well as VBM and CBM, were applied to all participants with group as the labels.

7.2.1 Experiment 1: Effect of Acquisition Site

The first experiment was to demonstrate the ability of SWPCA to reduce the intensity variance related to acquisition site. To do so, we first performed a VBM analysis in all five modalities (qT1, qT2, simulated T1 - weighted Inversion Recovery (synT1), GM and WM) separately, with the uncorrected (without applying SWPCA) and the corrected (after applying SWPCA) maps, using the acquisition site as labels.

To illustrate where the sources of variance of the acquisition sites are located, Figure 7.4 shows a brain t-map of significant ($p < 0.01$, $|t| > 2.57$) GM and WM between-site differences. The biggest reductions in variance were found in qT1 and synT1 maps, where high variability between acquisition sites, especially in the right hemisphere, was substantially reduced after the application of SWPCA. The reduction in the qT2, GM and WM maps was smaller, although noticeable.

To quantify the impact of this variance reduction on the between-groups effects, the classification analysis was undertaken. Higher accuracy values imply that the maps contain site-related patterns that were significant, whereas accuracy close to 0.5 indicates that the site-related variance was low. The test was applied to ALL, and also to the ASD and CTL subgroups. The classification results are presented in Table 7.1.

Performance results indicate clear advantages of using SWPCA, in particular in the case of qT1 and synT1 which were associated with strong site-dependent variance. These results are also consistent with the reduction of significant between-group areas observed in Figure 7.4.

The between-site differences were smaller for GM and WM maps, possibly due their reduced sensitivity. Since fractional occupancy values are abstract, unit-

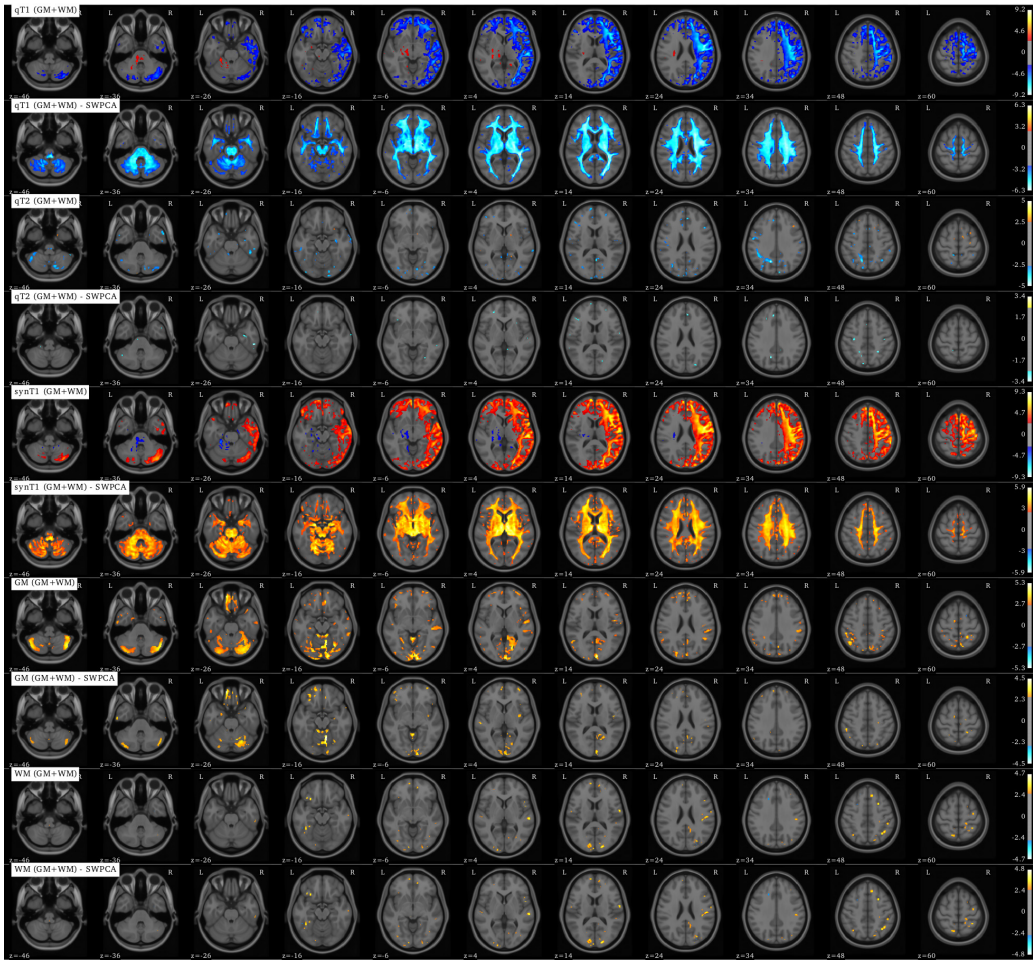


Figure 7.4: Brain t-map (VBM) of significant ($p < 0.01$, $|t| > 2.57$) GM and WM between-group differences using qT1, qT2, synT1, GM and WM modalities after applying SWPCA to remove site effects.

less values derived from each image they are less influenced by the acquisition site effects. For qT2 maps, the site-related differences were greater for the CTL participants than ASD where, according to the classification accuracy, they were nearly indistinguishable. Acquisition site differences were therefore noticeably reduced in the CTL and ALL databases, but not in the ASD.

7.2.2 Experiment 2: Within-site Between-Group Differences

In this second experiment, accuracy, sensitivity and specificity in the between-group comparison were recorded for images acquired from each site. This is an estimation of the discrimination ability of the different modalities without the influence of the site effects; Table 7.2. For all modalities, most of the values are close to a random classifier (~50%), indicative of having either no significant differences between groups, or having spatially heterogeneous patterns of sMRI measures across individuals where mass-univariate approaches are sub-optimal in detecting group differences. It is interesting to note that the London sample contained more between-group differences that those acquired in Cambridge.

7.2.3 Experiment 3: Effect of SWPCA on Group Differences

Finally, group differences were characterised with and without applying site-effects reduction via SWPCA to the five modalities.

Whole-brain VBM analysis was performed on the corrected and uncorrected maps from each modality. Figure 4 depicts the brain t-maps of significant ($p < 0.01$, $|t| > 2.57$) qT1, qT2, synT1, GM and WM between-group differences, using ALL, with the GM+WM mask, before and after applying SWPCA, so that the reduction of site-related variability can be observed. Some of the highlighted areas after applying SWPCA are inconsistent across modalities, with spurious peaks and noise, including a large area around the ventricles in the qT1 and synT1 modalities related to some abnormal participants that will be discussed later. However, there were some areas that were consistent across modalities. Significant areas found across at least 4 of the 5 modalities correspond to the Advanced Automated Labelling (AAL) (Tzourio-Mazoyer, et al., 2002) areas of: A) right superior frontal gyrus, Brodmann areas 6 (z=60); B) the pars opercularis of the left inferior frontal gyrus, Brodmann areas 44; C) the pars triangularis of the left inferior frontal gyrus, Brodmann areas 45; D) the posterior part of the left middle temporal gyrus (z=24); CSF filled spaces on the margins of the ventricles (z=-6,4,14,24); and the left crus I of cerebellar hemisphere (z=-26).

Modality	Mask	ALL		CTL		ASD	
		no-SWPCA	SWPCA	no-SWPCA	SWPCA	no-SWPCA	SWPCA
qT1	GM+WM	0.875 ± 0.083	0.530 ± 0.130	0.847 ± 0.141	0.543 ± 0.115	0.769 ± 0.145	0.553 ± 0.093
	GM	0.849 ± 0.085	0.535 ± 0.107	0.835 ± 0.154	0.501 ± 0.090	0.712 ± 0.161	0.575 ± 0.084
	WM	0.865 ± 0.082	0.447 ± 0.071	0.876 ± 0.128	0.441 ± 0.058	0.813 ± 0.127	0.575 ± 0.153
qT2	GM+WM	0.596 ± 0.128	0.503 ± 0.093	0.615 ± 0.196	0.454 ± 0.075	0.506 ± 0.192	0.476 ± 0.103
	GM	0.596 ± 0.126	0.493 ± 0.097	0.549 ± 0.187	0.478 ± 0.108	0.497 ± 0.197	0.425 ± 0.091
	WM	0.612 ± 0.131	0.560 ± 0.128	0.576 ± 0.195	0.550 ± 0.146	0.541 ± 0.185	0.575 ± 0.172
synT1	GM+WM	0.904 ± 0.073	0.563 ± 0.060	0.919 ± 0.100	0.440 ± 0.057	0.807 ± 0.151	0.631 ± 0.098
	GM	0.879 ± 0.090	0.576 ± 0.035	0.899 ± 0.108	0.526 ± 0.079	0.800 ± 0.145	0.587 ± 0.042
	WM	0.904 ± 0.076	0.582 ± 0.047	0.894 ± 0.111	0.574 ± 0.038	0.859 ± 0.112	0.468 ± 0.101
GM	GM+WM	0.595 ± 0.133	0.586 ± 0.141	0.582 ± 0.192	0.566 ± 0.093	0.481 ± 0.169	0.468 ± 0.152
	GM	0.620 ± 0.141	0.585 ± 0.078	0.604 ± 0.227	0.574 ± 0.038	0.499 ± 0.188	0.525 ± 0.114
WM	GM+WM	0.659 ± 0.139	0.448 ± 0.066	0.635 ± 0.180	0.507 ± 0.144	0.522 ± 0.206	0.525 ± 0.198
	WM	0.639 ± 0.124	0.549 ± 0.072	0.578 ± 0.194	0.516 ± 0.126	0.549 ± 0.160	0.526 ± 0.136

Table 7.1: Between-site classification accuracy ($\pm$ standard deviation) for different modalities and masks without and with SWPCA correction.

Modality	Mask	LONDON			CAMBRIDGE		
		acc.	sens.	spec.	acc.	sens.	spec.
qT1	GM+WM	0.603 ± 0.175	0.512 ± 0.260	0.692 ± 0.237	0.504 ± 0.193	0.492 ± 0.276	0.515 ± 0.307
	GM	0.501 ± 0.157	0.440 ± 0.244	0.565 ± 0.245	0.484 ± 0.201	0.488 ± 0.300	0.480 ± 0.327
	WM	0.505 ± 0.174	0.485 ± 0.248	0.526 ± 0.242	0.451 ± 0.197	0.465 ± 0.297	0.435 ± 0.296
qT2	GM+WM	0.628 ± 0.168	0.535 ± 0.246	0.719 ± 0.237	0.467 ± 0.181	0.527 ± 0.307	0.417 ± 0.314
	GM	0.539 ± 0.149	0.425 ± 0.220	0.654 ± 0.222	0.491 ± 0.196	0.548 ± 0.316	0.430 ± 0.298
	WM	0.619 ± 0.194	0.585 ± 0.262	0.655 ± 0.250	0.472 ± 0.195	0.448 ± 0.283	0.492 ± 0.290
synT1	GM+WM	0.665 ± 0.158	0.578 ± 0.224	0.755 ± 0.238	0.479 ± 0.201	0.478 ± 0.318	0.475 ± 0.316
	GM	0.547 ± 0.159	0.475 ± 0.237	0.622 ± 0.252	0.514 ± 0.218	0.477 ± 0.322	0.555 ± 0.342
	WM	0.515 ± 0.185	0.520 ± 0.288	0.506 ± 0.254	0.509 ± 0.209	0.472 ± 0.317	0.542 ± 0.316
GM	GM+WM	0.513 ± 0.171	0.507 ± 0.252	0.518 ± 0.245	0.488 ± 0.202	0.445 ± 0.318	0.528 ± 0.285
	GM	0.586 ± 0.174	0.610 ± 0.247	0.564 ± 0.270	0.521 ± 0.187	0.522 ± 0.303	0.535 ± 0.289
WM	GM+WM	0.471 ± 0.181	0.455 ± 0.245	0.488 ± 0.278	0.489 ± 0.206	0.502 ± 0.319	0.483 ± 0.314
	WM	0.465 ± 0.174	0.445 ± 0.243	0.484 ± 0.268	0.468 ± 0.210	0.488 ± 0.292	0.448 ± 0.305

Table 7.2: Classification accuracy (Acc), sensitivity (Sen) and specificity (Spec) $\pm$ standard deviation for each modality and mask using the participants acquired at the LON and CAM sites.

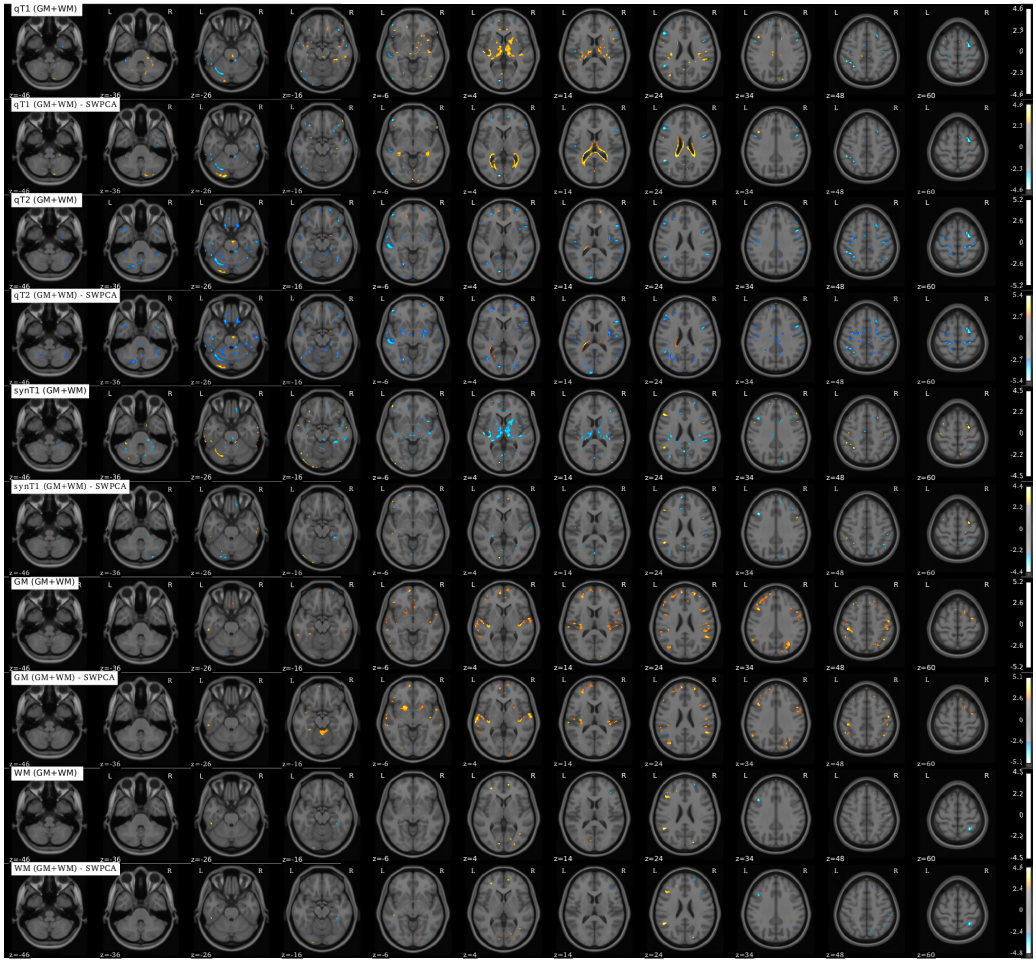


Figure 7.5: Brain t-map (VBM) of significant ($p < 0.01$, $|t| > 2.57$) GM and WM differences in ASD using qT1, qT2, synT1, GM and WM maps before and after applying SWPCA to remove site effects.

The complementary CBM (Section 2.4) analysis was performed on the most significant components. The resulting regions, statistically thresholded with $Z > 2.57$ (corresponding to $p < 0.01$), were superimposed on the MNI template, and are depicted in Figure 7.6. A reduction of significant between-group areas after applying SWPCA is evident in most modalities, but particularly noticeable in the qT1 and qT2. In WM no significant regions were observed, neither before nor after SWPCA. The significant regions identified in any modality corresponded to the AAL areas of the CSF filled areas around the ventricles (planes z=-6, 4, 14, 24), the right middle temporal gyrus (plane z=14) and the left crus I of cerebellar hemisphere (plane z=-26). However, none of these regions were repeated over more than two of the modalities, except for the large areas around ventricles that were caused by abnormalities in three participants, which will be discussed later.

Performance results for the classification analysis applied to ALL are shown in Table 7.3. Between-group results were quite similar before or after applying SWPCA, although reducing between-site variance generally reduced the performance towards a random classifier. The results in this table match the overall effects that were found in Figure 4, where most spurious significance peaks disappeared after applying SWPCA, but some regions were highlighted. These regions, where SWPCA did not seem to eliminate the significant areas but enhanced them, could be responsible for the accuracy increment in the analysis of the qT2 modality, and the GM with GM mask.

7.2.4 Discussion

Brain anatomical and functional differences between ASD participants and controls have been explored by a number of previous studies (Di Martino, et al., 2014; Ecker, et al., 2015; Hernandez, et al., 2015; Lenroot and Yeung, 2013; Zürcher, et al., 2015). Many affected structures have been proposed in each of these studies, however as a recent large-scale study points out (Haar, et al., 2014), these are frequently inconsistent throughout the literature. Researchers argue that most of these structures are database-dependent, and since many studies use multi-site acquisition procedures, the variance introduced by each acquisition site is a probable source of Type I errors.

The technical and logistical drawbacks of multicentre studies are widely documented, including participant recruitment procedures (Pearlson, 2009) and technical effects that range from the usage of different equipment or acquisition parameters (Van Horn and Toga, 2009) to physical changes that affect the performance of MRI scanners across time (Pearlson, 2009). There is general recognition

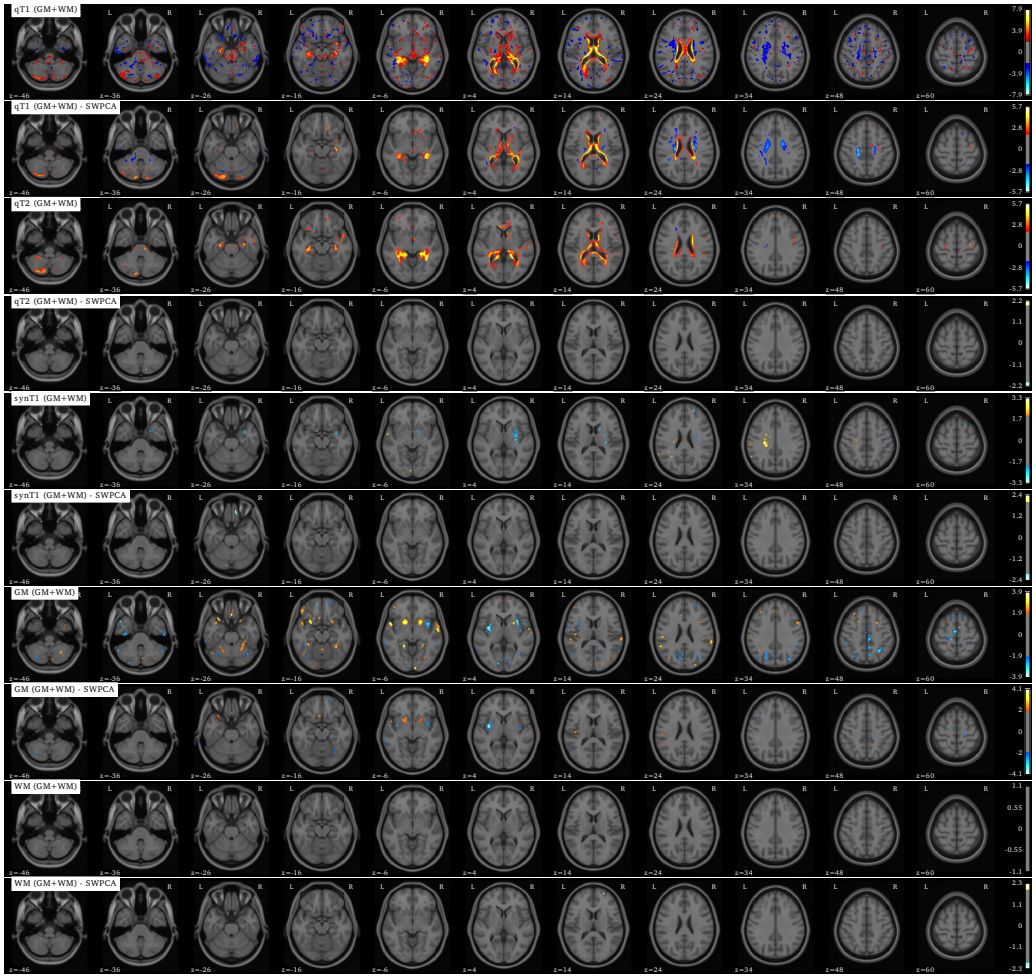


Figure 7.6: Brain Z-map (CBM) of significant ($p < 0.01$, $|t| > 2.57$) GM and WM differences using qT1, qT2, synT1, GM and WM maps before and after applying SWPCA to remove site effects.

Modality	Mask	NO-SWPCA			SWPCA		
		acc.	sens.	spec.	acc.	sens.	spec.
qT1	GM+WM	0.564 ± 0.123	0.503 ± 0.179	0.625 ± 0.177	0.435 ± 0.123	0.499 ± 0.181	0.371 ± 0.178
	GM	0.523 ± 0.112	0.468 ± 0.162	0.580 ± 0.192	0.458 ± 0.120	0.477 ± 0.187	0.441 ± 0.210
	WM	0.504 ± 0.131	0.475 ± 0.191	0.533 ± 0.194	0.484 ± 0.123	0.511 ± 0.179	0.456 ± 0.194
qT2	GM+WM	0.578 ± 0.115	0.487 ± 0.208	0.669 ± 0.178	0.593 ± 0.136	0.546 ± 0.206	0.640 ± 0.194
	GM	0.554 ± 0.135	0.492 ± 0.194	0.614 ± 0.181	0.526 ± 0.144	0.512 ± 0.209	0.543 ± 0.222
	WM	0.516 ± 0.138	0.508 ± 0.198	0.522 ± 0.216	0.499 ± 0.137	0.477 ± 0.209	0.521 ± 0.196
synT1	GM+WM	0.596 ± 0.132	0.509 ± 0.194	0.680 ± 0.172	0.577 ± 0.130	0.479 ± 0.208	0.676 ± 0.183
	GM	0.587 ± 0.139	0.509 ± 0.210	0.665 ± 0.169	0.483 ± 0.136	0.489 ± 0.218	0.480 ± 0.200
	WM	0.496 ± 0.139	0.500 ± 0.189	0.492 ± 0.194	0.487 ± 0.134	0.513 ± 0.189	0.461 ± 0.211
GM	GM+WM	0.498 ± 0.120	0.486 ± 0.197	0.507 ± 0.203	0.490 ± 0.123	0.514 ± 0.197	0.465 ± 0.182
	GM	0.574 ± 0.121	0.571 ± 0.189	0.579 ± 0.163	0.593 ± 0.127	0.602 ± 0.172	0.587 ± 0.190
WM	GM+WM	0.499 ± 0.132	0.506 ± 0.189	0.487 ± 0.181	0.521 ± 0.129	0.510 ± 0.209	0.532 ± 0.180
	WM	0.506 ± 0.143	0.488 ± 0.219	0.526 ± 0.197	0.507 ± 0.122	0.521 ± 0.165	0.492 ± 0.193

Table 7.3: Classification accuracy (Acc), sensitivity (Sen), and specificity (Spec) $\pm$ STD for the different modalities and masks using ALL, before and after applying SWPCA.

that standardization is needed to ensure the uniformity of the acquired maps. Different approaches have been used in large-scale studies, such as ADNI where human "phantoms" were used to perform a preparatory optimisation of MRI scanning platforms (Friedman and Glover, 2006).

There are two major types of site effects, regardless of their source: geometric distortions and intensity inhomogeneities. In this work, we focused on the latter, since much of the geometric distortion has been eliminated during acquisition (see Section 2.1), and the DARTEL normalization and registration acts as a homogenizing step, reducing both between-site and between-subject geometric differences, substantially reducing the impact of the site-related geometric differences.

Regarding intensity correction, in the MRC AIMS database used in this study (Ecker, et al., 2013; Ecker, et al., 2012), a standardization procedure based on quantitative imaging (Deoni, et al., 2008) was used to minimize inter-site variance and improve the signal-to-noise contrast. However, as the between-site analysis in Section 7.2.1 suggests, this strategy still results in variance that makes it easier to distinguish scanning sites than diagnostic groups. For example, when using qT1 the accuracy for LON vs. CAM classification was >80%, whilst when classifying ASD vs. CTL it was 52%. This marks the substantial effect of site variance on the maps' intensity distribution, even when the multi-site study employs quantitative imaging protocol on the same model of scanner platform across sites. However, with the inclusion of GM and WM maps, we can observe that the inhomogeneities found on qT1 or synT1 barely affected the segmentation procedure.

In this work, the approach we have taken is to perform a multivariate decomposition of each dataset into a number of components that explain different portions of variance. The following step was to identify the components of variance that are due to multi-site acquisition and reduce them. Decomposition was completed using PCA and then, to identify which of the components were linked to acquisition site, we performed an ANOVA on the component scores. Finally, using the weighting function defined in Sec. 7.2.3, we reconstructed the original signal reducing the undesired variance, in what we called Significance Weighted PCA (SWPCA). The method has proven its ability in reducing undesired variance, quantifiable by means of the accuracy obtained in a site vs. site classification. In this case, SWPCA reduced the accuracy from >0.8 to approximately ˜0.5, a random classifier, suggesting that most site-related variance was eliminated.

A simpler approach such as applying a voxel-by-voxel ANOVA would also be useful to reduce the acquisition site effects (Suckling, et al., 2012). However, SWPCA is a multivariate approach that still offers major advantages over this

voxel-wise algorithm, and similar algorithms have found utility in text document searches (Kriegel, et al., 2008; Tavoli, et al., 2013; Zhang and Nguyen, 2005). First, PCA models the different sources of variance of the dataset, whereas a simple voxel-wise ANOVA only removes mean site differences, which might result in less statistical power. Secondly, SWPCA is multivariate in nature, where each component contains information that potentially affects all voxels. Together, these two features allow SWPCA to identify the components linked to the undesired effects, and reduce their impact with a weighted reconstruction approach, reducing the general variance related to the acquisition site. However, this increased power reveals a major drawback: SWPCA needs at least a moderate number of participants to work properly. That is the reason why we cannot apply SWPCA to databases such as ADNI (Friedman and Glover, 2006) or ABIDE (Di Martino, et al., 2014), where the number of participants acquired at each site is small, or to the six travelling phantoms used in the calibration of the MRC AIMS study.

There exist a number of similar multivariate methods that model the influence of categorical variables, such as the well-known Partial Least Squares (PLS) algorithm (Vinzi, et al., 2010) or Surrogate Variable Analysis (SVA) (Leek and Storey, 2007). In the first case, both PLS and SWPCA take categorical variables $\mathbf{Y}$ along with the data $\mathbf{X}$ as inputs to partition the influence of these into components. However, the most significant difference is the underlying model. Whilst SWPCA estimates the principal components blindly using their variance, which is what we aim to reduce, and performs an ANOVA afterwards, PLS uses the categorical variable in the computation of the covariance matrix and then estimates the components.

On the other hand, SVA, used for gene expression studies (Leek and Storey, 2007), is more comparable to SWPCA. The SVA algorithm uses a number of decomposition and significance estimation steps to construct a set of surrogate variables; that is, variables that account for the unmodeled variance and expression heterogeneity. While similar to SWPCA in the steps used (i.e. SVD decomposition and significance estimation), their approaches are fundamentally different. SVA constructs a higher complexity model that starts by eliminating the contribution of primary variables to produce a number of unknown hidden (surrogate) variables, whereas SWPCA is intended to reduce complexity by producing variance-reduced maps to reduce the influence of previously known, but unconsidered, variables and facilitate a subsequent analysis focused only on the relevant variables.

Focusing on the VBM results, after performing the site-effects removal by SWPCA significant between-group differences were noted in five areas: A) the right superior frontal gyrus; B) the pars opercularis of the left inferior frontal gyrus;

C) the pars triangularis of the left inferior frontal gyrus; D) the posterior part of the left middle temporal gyrus; and E) the left crus I of cerebellar hemisphere. The first three regions are within Brodmann areas 6, 44 and 45. However, when examining the projection of the region D onto the MNI template (see Figure 6), it is also located in the posterior part of the left superior temporal gyrus. Therefore, D corresponds closely with the region between Brodmann areas 22 and 39, the Temporo-Parietal Junction (TPJ), with negative t-value at the left side (containing Wernicke's area) and positive t-value at the right side.

The role of these regions in autism has received much attention. Brodmann areas 44 and 45, that together make the Broca's Area (of importance in speech production and a proposed part of the human mirror neuron system (Nishitani, et al., 2005)), is a region where mirror neuron dysfunction has been consistently reported in ASD-affected children (Dapretto, et al., 2006) and adults (Hadjikhani, et al., 2006; Lopez-Hurtado and Prieto, 2008; Verly, et al., 2014). Wernicke's area, contained in the left TPJ, is also linked to language, and has been associated with ASD in several works (Hadjikhani, et al., 2006; Kriegel, et al., 2008; Verly, et al., 2014). Additionally, the right TPJ has been proposed as related to mentalizing and has been repeatedly implicated in autism (Barnea-Goraly, et al., 2004), including a fMRI study of a subsample of this same AIMS dataset (Lombardo, et al., 2011). The right superior frontal gyrus (region A) is more equivocal, with some studies (Ecker, et al., 2010; Ecker, et al., 2012) reporting abnormalities in this area, while others (Hadjikhani, et al., 2006; Segovia, et al., 2014) report no significant differences. Our analyses reveal no differences in the insula and amygdala, brain structures frequently linked to autism.

Some regions, particularly in qT2, synT1 and segmented GM maps show potentially spurious significance peaks around the ventricles and especially in the left crus I of cerebellar hemisphere (region E). After examining the database, two individuals had appreciable structural abnormalities in the form of abnormal ventricle size and cerebellar atrophy, as can be seen in Figure 7. It is possible that these participants influenced the computation of the *t*-maps, and therefore are responsible for the significance in region E and areas surrounding the ventricles and, since they are part of the LON subdataset, could also be responsible for the increased classification accuracy of the quantitative T1 and T2, and the synthetic T1 maps in this sub-dataset.

After observing the influence of these participants on the computation of the *t*-maps, we can assume that most of the structural differences in ASD are so subtle that the influence of just one or two images can impact on the final results. This, along with the poor performance of the classification pipeline presented in Section 3, dramatically reduces the significance of the aforementioned *t*-maps. Therefore, the existing evidence leads to the conclusion that ASD presents as ei-

ther undetectable structural differences or, more likely, with such heterogeneous differences that are difficult to establish a common pattern even after reducing the variance introduced by acquisition site.

It may be the case that cohorts of individuals examined at different sites are somehow systematically biased towards a specific type of patient (in ways that we cannot see simply based on phenotypic information), then site-related intensity variability is also enriched with important variability about nested autism subgroups. So with any technique trying to remove the site-related inhomogeneity, the subgroup information could also be removed. Together, the evidence supports the claim that defining meaningful subgroups based on different measures, such as genetic profiling, clinical co-morbidities or sensory sensitivities, is the most urgent next step for ASD research (Haar, et al., 2014).

7.3 Results for DaTSCAN Datasets

		Performance		
SWPCA	Norm.	acc.	sens.	spec.
no	max	0.883 ± 0.030	0.855 ± 0.058	0.915 ± 0.058
	int	0.877 ± 0.035	0.849 ± 0.073	0.908 ± 0.079
	stable	0.898 ± 0.033	0.883 ± 0.057	0.915 ± 0.079
yes	max	0.539 ± 0.100	0.527 ± 0.373	0.550 ± 0.337
	int	$\pm$	$\pm$	$\pm$
	stable	0.361 ± 0.102	0.394 ± 0.295	0.322 ± 0.270

Table 7.4: Performance measures for the combined DaTSCAN dataset found before and after applying SWPCA.

8 SIMULATION OF FUNCTIONAL BRAIN IMAGES

8.1 Simulation Procedure

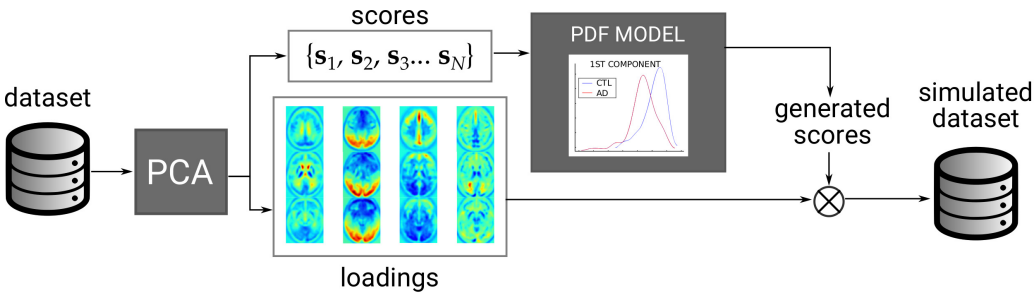


Figure 8.1: Schema of the brain image synthesis algorithm.

8.1.1 Decomposition via PCA

The first step in our simulation algorithm is to project the original dataset to a new space defined by the principal components of the set; that is, the eigenbrain space. In this space, each subject from the original dataset is projected to a point, and we can afterwards use the space basis (the principal components) to reconstruct that particular subject.in this work we will use the first N components for performance, where N is the number of subjects that are used in the computation of PCA. For more details about PCA, see Section 7.1.1.

8.1.2 Probability Density Modelling using Kernel Density Estimation

Kernel Density Estimation (KDE) is used here to model the statistical distribution of the projected subjects in the eigenbrain space, and it is applied independently to each AD, MCI and CTL class. The KDE estimates the probability density

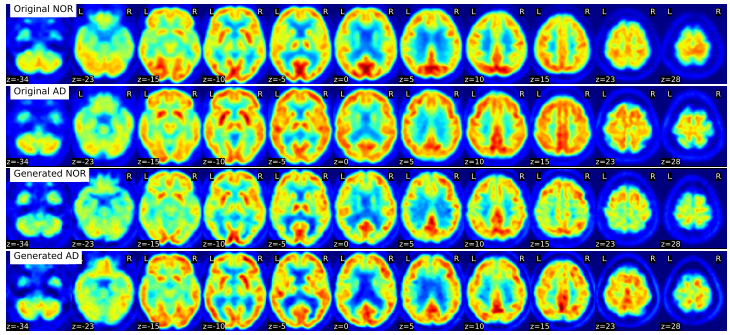


Figure 8.2: Comparison between simulated and original images from AD and CTL classes.

function f from a number of independent and identically distributed samples $(x_1, x_2, \ldots x_n)$, in the following manner:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big), \tag{8.1}$$

where $h > 0$ is the bandwidth, a smoothing parameter. The KDE via diffusion **[Botev2010]** used in this article uses a data-driven automatic estimation of the bandwidth, which unlike most methods, does not rely on arbitrary normal reference rules.

8.1.3 Probability Density Modelling using Multivariate Gaussian

8.1.4 Random Number Generation

8.1.5 Brain Image Synthesis

8.2 Experimental Setup

To validate the simulated dataset, we have performed two different experiments:

- **Exp. 1**: We have estimated the predictive power of the simulated images by generating new images from the original training set in each cross-validation iteration and using them to predict the original test set.
- **Exp. 2**: We tested that the simulated images are independent from the original ones, although preserving similar characteristics. To do so, following a Voxel as Features (VAF) approach **[Stoeckel04]**, we extract a small subset

(10 AD and 10 NOR) from the original dataset. Then, we trick the classifier, training it with the whole subset -instead of the training set only-, and testing it against the test set. Therefore, the performance of the tricked system must be close to 1. Then, we generate a new set of simulated images (100 AD and 100 NOR) from the reduced subset and proceed similarly. If our simulated images are independent from the originals, the performance of the system should decrease substantially.

Classification is performed using a Support Vector Machine (SVM) classifier with linear kernel. Estimation of parameter C is performed in an inner cross-validation loop within the training set. Values of accuracy (acc), sensitivity (sens) and specificity (spec) and their standard deviation (SD) are estimated.

8.3 Results for ADNI-PET Dataset

8.3.1 Experiment 1

The performance results for the proposed experiments are shown in Table 8.1. Exp. 1 is applied to three different scenarios: only AD vs NOR (95 vs 101 subjects), and after incorporating MCI subjects, using them as NOR or AD.

Scenario	acc ($\pm$SD)	sens ($\pm$SD)	spec ($\pm$SD)
AD vs NOR	0.882 ± 0.012	0.865 ± 0.091	0.901 ± 0.118
MCI as NOR	0.727 ± 0.119	0.769 ± 0.155	0.789 ± 0.151
MCI as AD	0.739 ± 0.126	0.747 ± 0.147	0.845 ± 0.146

Table 8.1: Baseline performance of the set, using the original dataset.

Scenario	acc ($\pm$SD)	sens ($\pm$SD)	spec ($\pm$SD)
AD vs NOR	0.801 ± 0.095	0.782 ± 0.202	0.821 ± 0.191
MCI as NOR	0.751 ± 0.078	0.433 ± 0.201	0.851 ± 0.262
MCI as AD	0.712 ± 0.048	0.821 ± 0.062	0.382 ± 0.248

Table 8.2: Performance of Exp 1, demonstrating the predictive ability of the simulated images over the real dataset.

Scenario	acc (±SD)	sens (±SD)	spec (±SD)
Original	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Simulated	0.839 ± 0.094	0.830 ± 0.228	0.849 ± 0.206

Table 8.3: Performance of the Exp 3 proves the independence of the simulated images with respect to the originals.

8.4 Results for DaTSCAN Datasets

Part IV

GENERAL DISCUSSION AND CONCLUSIONS

9 GENERAL DISCUSSION AND CONCLUSIONS

9.1 General Discussion

9.1.1 Discussion on the algorithms

9.1.2 Discussion on the diseases

9.2 Conclusions

9.3 Future Work

asdf

Part V

APPENDIX

A DATASETS

Many dataset are used in this thesis, covering three imaging modalities and three disorders. A summary of these can be found on Table A.1, folowed by a longer description of each one.

Acronym	Origin	Disease	Modality	Drug
ADNI-MRI	ADNI	AD	MRI	-
AIMS-MRI	MRC-AIMS	ASD	MRI	-
ADNI-PET	ADNI	AD	PET	**HMPAO!**
VDLN-HMPAO	VDLN	AD	SPECT	**HMPAO!**
VDLN-DAT	VDLN	PKS	SPECT	DaTSCAN
VDLV-DAT	VDLV	PKS	SPECT	DaTSCAN
PPMI-DAT	PPMI	PKS	SPECT	DaTSCAN

Table A.1: Summary of the datasets used in this thesis.

A.1 Magnetic Resonance Imaging

A.1.1 ADNI-MRI, Alzheimer's Disease Neuroimaging Initiative

AD

A.1.2 AIMS-MRI, MRC-AIMS Consortium

Structural MRI were analysed from 136 adult, right-handed males (68 with ASD and 68 matched controls) with no significant mean differences in age and full-scale IQ, acquired from the centres contributing to the UK Medical Research Council Autism Imaging Multi-centre Study (MRC AIMS) (Ecker, et al., 2013; Ecker, et al., 2012) and recruited by advertisement. In this work, only participants recruited at the Institute of Psychiatry, King's College London (LON) and

the Autism Research Centre, University of Cambridge (CAM) were included where an equivalent set of images were acquired from each participant.

Participants were excluded from the study if they had a history of major psychiatric disorder or medical illness affecting brain function (e.g. psychosis or epilepsy), or current drug misuse (including alcohol), or were taking antipsychotic medication, mood stabilizers or benzodiazepines.

All participants with ASD were diagnosed according to International Classification of Diseases, 10th Revision (ICD-10) research criteria, and confirmed using the Autism Diagnostic Interview-Revised (ADI-R) (Lord, et al., 1994). Autism Diagnostic Observation Schedule (ADOS) (Lord, et al., 2000) was performed, but the score was not considered as an inclusion criteria. ASD participants, to be included, must have scored above the ADI-R cut-off in the three domains of impaired reciprocal social interaction, communication and repetitive behaviours and stereotyped patterns, although failure to reach cut-off in one of the domains by one point was permitted. Intellectual ability was assessed using the Wechsler Abbreviated Scale of Intelligence (WASI) (Wechsler, 1999), ensuring the participants fell within the high-functioning range on the spectrum defined by a full-scale IQ > 70. The demographics of the participants are shown in detail in Table A.2.

Database	Group	N	Age ($\mu \pm \sigma$ years)	IQ ($\mu \pm \sigma$)
LON	ASD	39	28.74 ± 6.52	111.28 ± 13.13
	CTL	40	25.30 ± 6.62	104.67 ± 11.16
CAM	ASD	29	26.83 ± 4.64	115.83 ± 11.88
	CTL	28	26.75 ± 7.32	115.25 ± 13.67
ALL	ASD	68	25.90 ± 6.95	109.03 ± 13.31
	CTL	68	27.93 ± 5.87	113.22 ± 12.81

Table A.2: Demographics of the AIMS-MRI dataset.

Structural MRI were obtained using Driven Equilibrium Single Pulse Observation of T1 and T2 (DESPOT1, DESPOT2) (Deoni, et al., 2008) at King's College London and University of Cambridge, both with 3T GE Medical Systems HDx scanners. Using multiple Spoilt Gradient Recall (SPGR) acquisitions in the DESPOT1 sequence and Steady State Free Procession (SSPF) acquisitions in the DESPOT2 sequence, with different flip angles and repetition times, qT1 and qT2 maps were calculated with a custom ImageJ plug-in package. Correction of main and transmit magnetic field (B0 and B1) inhomogeneity effects was performed during the estimation of T1 and T2.

For accurate registration to the standard stereotatic space of the MNI, a synT1 images were created based on the qT1 maps (Ecker, et al., 2013; Ecker, et al., 2012; Lai, et al., 2012). The synT1 images were then segmented using New Segment into GM and WM maps, and normalized to the MNI space using DARTEL in SPM8 (Friston, et al., 2007), with modulation (preserve volume) to retain information of regional/local GM and WM volumes, and smoothed with a 3mm FWHM Gaussian Kernel to account for inter-subject mis-registration. The synT1, qT1 and qT2 maps were also registered to the standard MNI space using the same DARTEL flow fields, but without modulation (preserve concentration) to retain information of regional/local T1 contrast, T1 relaxation time, and T2 relaxation times, and smoothed with a 3mm FWHM Gaussian kernel. Therefore, there were five different modalities: qT1, qT2, synT1 map, GM and WM maps, for each every participant, which allows us to observe the impact of our SWPCA correction of site-related undesired variance on quantitative (qT1 and qT2), simulated (synT1) images and probability maps (GM and WM).

During the pre-processing of the images, several procedures targeted the reduction of inter-subject and inter-site geometric distortion, amongst them the correction of B0 and B1 field inhomogeneity effects and the registration to MNI space. Many other algorithms have been proposed to help in this task. However, the study of their relative performance lies beyond the scope of this article. Following image registration, it was assumed that only the intensity of the maps was affected between sites.

A.2 Positron Emission Tomography

A.2.1 ADNI-PET, Alzheimer's Disease Neuroimaging Initiative

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and AD. For up-to-date information, see `www.adni-info.org`.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). In this work, the ^{18}F-FDG PET images, used to estimate the metabolic activity of the brain, are used to generate and validate the simulated images. 95 PET images from AD affected subjects, 207 images from Mild Cognitive Impairment (MCI)

affected subjects and 101 images from Normal Controls (NOR) have been used to construct the original N = 403 set from which the simulation parameters will be obtained.

A.3 Single Photon Emission Computed Tomography

A.3.1 VDLN-HMPAO, Virgen de las Nieves

The database is built up of imaging studies of subjects following the protocol of an hospital-based service. First, the neurologist evaluated the cognitive function, and those patients with findings of memory loss or dementia were referred to the nuclear medicine department in the "Virgen de las Nieves" hospital (Granada, Spain), in order to acquire complementary screening information for diagnosis[1]. Experienced physicians evaluated the images visually. The images were assessed using 4 different labels: CTL for subjects without scintigraphic abnormalities and mild perfusion deficit (AD1), moderate deficit (AD2) and severe deficit (AD3), to distinguish between different levels of presence of hypo-perfusion patterns compatible with AD. In total, the database consists of n =97 subjects: 41 CTRL, 30 AD1, 22 AD2 and 4 AD3 (see table A.3 for demographic details). Since the patients are not pathologically confirmed, the subject's labels possesses some degree of uncertainty, as the pattern of hypo-perfusion may not reflect the underlying pathology of AD, nor the different classification of scans necessarily reflect the severity of the patients symptoms. However, when pathological information is available, visual assessments by experts have been shown to be very sensitive and specific labelling methods, in contrast to neuropsychological tests [**jobst_accurate_1998**, **dougall_systematic_2004**]. Given that this is an inherent limitation of 'in vivo' studies, our working-assumption is that the labels are true, considering the subject label positive when belonging to any of the AD classes, and negative otherwise.

A.3.2 VDLN-DAT, Virgen de las Nieves

SPECT DATSCAN

73 CTL, 45 PD, 30 SWEDD.

1 Clinical information is unfortunately not available for privacy reasons, but only demographic information

	#samples	Sex(M/F)(%)	μ[range/σ]
CTRL	41	32.95/12.19	71.51[46-85/7.99]
AD1	29	10.97/18.29	65.29[23-81/13.36]
AD2	22	13.41/9.76	65.73[46-86/8.25]
AD3	4	0/2.43	76[69-83/9.90]

Table A.3: Demographic details of the ADNI-PET dataset. CTRL = Normal Controls, AD 1 = possible AD, AD 2 = probable AD, AD 3 = certain AD. μ and σ stands for population mean and standard deviation respectively.

A.3.3 VDLV-DAT, Virgen de la Victoria Hospital

The images were obtained after a period of between 3 and 4 hours after the intravenous injection of 185 MBq (5 mCi) of DaTSCAN, with prior thyroid blocking with Lugol's solution. The tomographic study (SPECT) with Ioflupane/FP-CIT-I-123 was performed using a General Electric gamma camera, Millennium model, equipped with a dual head and general purpose collimator. A 360-degree circular orbit was made around the cranium, at 3-degree intervals, 60 images with a duration of 35 seconds per interval, 128×128 matrix. Image reconstruction was carried out using filtered back-projection algorithms without attenuation correction [**Shepp82**, **Vardi1985**], application of a Hanning filter (frequency 0.7) and images were obtained with transaxial cuts, following the method proposed in [**Ramirez2009**].

The images were interpreted by three Nuclear Medicine specialists, with masking of the clinical orientation. Visual assessment was established by exclusively considering the normal/abnormal criterion and after arriving at a consensus report between the three specialists, i.e. whether the FP-CIT SPECT allowed differentiation of a group of conditions with presynaptic involvement from others in which their integrity is assumed, without trying to assign them to different clinical groups within the set of pathological studies. A study was considered to be normal when bilateral, symmetrical uptake appeared in caudate and putamen nuclei, and abnormal when there were areas of qualitatively reduced uptake in any of the striatal structures.

A total of 208 subjects (100 patients and 108 controls), randomly selected from the total studies performed in this center until December 2008 and referred to it because of a movement disorder, were included in the study. Mean age was 70.2 years (41-87) with a standard deviation of 10.2 years (a detailed description of the database can be found in [**Lozano2007**]). Clinical diagnosis, a parameter used as 'gold Standard' to establish the existence of PS, was made using the

diagnostic criteria established previously, with an established minimum follow-up period of 18 months. Those patients who were receiving treatment with drugs that had known or suspected effect on the level of the dopaminergic transporters through direct competitive mechanism were excluded. Although PD is the most representative pathology of PS, there are other medical conditions which, though they differ clinically from this, are also expressed by this set of symptoms. Some of them are multisystem atrophy (MSA), progressive supra-nuclear palsy (PSP) and corticobasal degeneration (CBD), in which, unlike PD, as well as involvement of the presynaptic terminal, there is involvement at the post-synaptic level of the nigrostriatal pathway.

A.3.4 PPMI-DAT, Parkinson's Progression Markers Initiative

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org.

The images in this database were imaged 4 + 0.5 hours after the injection of between 111 and 185 MBq of DaTSCAN. Subjects were also pretreated with saturated iodine solution (10 drops in water) or perchlorate (1000 mg) prior to the injection. All subjects had a supplied ^{57}Co line marker affixed along the canthomeatal line, which will facilitate subsequent image processing and allow the core lab to accurately distinguish left and right in the face of multiple image file transfers. These markers are only evident in the ^{57}Co window and hence do not contaminate the ^{123}I-DaTSCAN brain data [**PPMI**, **Inititative2010**].

111 CTL, 32 SWEDD and 158 PD

Raw projection data are acquired into a 128×128 matrix stepping each 3 degrees for a total of 120 projection into two 20% symmetric photopeak windows centered on 159 KeV and 122 KeV with a total scan duration of approximately 30 - 45 minutes. Other scan parameters (collimation, acquisition mode, etc.) are selected for each site. The images of both the subject's data and the cobalt striatal phantom are reconstructed and attenuation corrected, implementing either filtered back-projection or an iterative reconstruction algorithm using standardized approaches [**Inititative2010**]. After the processing, the database contains 289 spatially normalized images, 114 from Normal Control subjects and 175 from PD patients, and of a $91 \times 109 \times 91$ size.

B BACKGROUND ON SUPPORT VECTOR MACHINES

Support Vector Machine (SVM), introduced in the late 70s [**Vapnik1982**], are a set of related supervised learning methods widely used in pattern recognition, voice activity detection (VAD), classification and regression analysis.

We suppose the data to be linearly separable. In this case, a data point is viewed as a p-dimensional vector. Our objective is to separate a set of binary labelled training data with a hyperplane that is maximally distant from the two classes (known as the maximal margin hyper-plane). To do so, we build a function $f : \mathfrak{R}^n -> \{\pm 1\}$ using training data that is, p-dimensional patterns x_i and class labels y_i :

$$(\mathbf{x_1}, \mathbf{y_1}), (\mathbf{x_2}, \mathbf{y_2}), ..., (\mathbf{x_l}, \mathbf{y_l}) \in \mathfrak{R}^n \times \{\pm 1\} \tag{B.1}$$

so that f will correctly classify new examples (**x**, **y**).

Linear discriminant functions define decision hypersurfaces or hyperplanes in a multidimensional feature space, that is:

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \omega_0 = 0, \tag{B.2}$$

where **w** is known as the weight vector and ω_0 as the threshold. The weight vector **w** is orthogonal to the decision hyperplane and the optimization task consists of finding the unknown parameters $\omega_i, i = 1, ..., n$ defining the decision hyperplane.

Let $\mathbf{x_i}, i = 1, 2, ..., n$ be the feature vectors of the training set, X. These belong to either of the two classes, ω_1 or ω_2. If the classes were linearly separable, the objective would be to design a hyperplane that classifies correctly all the training vectors. The hyperplane is not unique, and the selection process is focused on maximizing the generalization performance of the classifier, that is, the ability of the classifier, designed using the training set, to operate satisfactorily with new data. Among the different design criteria, the maximal margin hyperplane is usually selected since it leaves the maximum margin of separation between the two classes. Since the distance from a point **x** to the hyperplane is given by $z = |g(\mathbf{x})|/\|\mathbf{w}\|$, scaling w and $\mathbf{w_0}$ so that the value of $g(\mathbf{x})$ is $+1$ for the nearest point in ω_1 and -1 for the nearest points in ω_2, reduces the optimization problem to maximizing the margin: $2/\|\mathbf{w}\|$ with the constraints:

$$\mathbf{w}^T\mathbf{x} + \mathbf{w_o} \geqslant 1, \forall \mathbf{x} \in \omega_1 \tag{B.3}$$

$$\mathbf{w}^T\mathbf{x} + \mathbf{w_o} \leqslant 1, \forall \mathbf{x} \in \omega_2 \tag{B.4}$$