Universidad Carlos III of Madrid

Degree in Audiovisual Systems Engineering

BACHELOR THESIS

Automatic Design of Neuromarkers for Obsessive Compulsive Disorder Characterisation

Author:

Óscar García Hinde

Tutor:

Dr. Vanessa Gómez Verdejo

Acknowledgements

My parents for their love and endless support.

My tutor, Dr. Vanessa Gómez Verdejo, for showing me an entire intellectual universe beyond my degree and trusting me enough to lead me through it.

Dr. Emilio Parrado Hernández, for bringing vision and context to my work and for offering unlimited amounts of patience.

Dr. Manel Martínez Ramón and Dr. Carles Soriano Mas for all the work they did previous to this thesis and for trusting a lowly undergrad like me to meddle with their research.

Ana. This is it! :)

Abstract

This bacherlor thesis proposes a new paradigm to discover biomarkers capable of characterizing obsessive-compulsive disorder (OCD) by means of machine learning methods. These biomarkers, named neuromarkers, will be obtained through the analysis of sets of magnetic resonance images of the brains of OCD patients and healthy control subjects.

The design of the neuromarkers stems from a method for the automatic discovery of clusters of voxels, distributed in separate brain regions, relevant to OCD. This method was recently published by Dr. Emilio Parrado Hernández, Dr. Vanessa Gómez Verdejo and Dr. Manel Martínez Ramón.

With these clusters as a starting point, we will define the neuromarkers as a set of measurements describing features of these individual regions. Then we will perform a selection of these neuromarkers, using state of the art feature selection techniques, to arrive at a reduced, relevant and intuitive set.

The results will be sent to Dr. Carles Soriano Mas at the Bellvitge University Hospital in Barcelona, Spain. His feedback will be used to determine the efficacy of our neuromarkers and their usefulness for psychiatric analysis.

The main goal of the project is to come up with a set of neuromarkers for OCD characterisation that are easy to interpret and handle by the psychiatric community.

A paper presenting the methods and results described in this bachelor thesis, of which the student is the main author, has been submitted and accepted for presentation in the 2014 European Congress of Machine Learning (ECML/PKDD 2014). The ECML reported a 23.8% paper acceptance rate for 2014.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem description

In some areas of medicine it is quite common to find punctuation systems that allow for state evaluation and patient diagnosis. For instance APACHE II (Acute Physiology and Chronic Health Evaluation) [30] is one of the most widely used score-based systems to quantify the seriousness of critical patient's state by means of 12 factors or routine physiological measures (blood pressure, body temperature, heart rate, etc.). Other important score-based systems are the Ranson criterion, which predicts the severity of acute pancreatitis [44], the Glasgow scale [27], used to measure a person's conscience level, or the SAPS II index (Simplified Acute Physiology Score) [32] which, as the APACHE II index, estimates the severity of a patient's state. It has been shown that the adequate use of these scores provides a better characterisation of the illness and helps researchers analyse the success of new therapies and compare their effectiveness in different hospitals.

However, psychiatry lacks direct and objective indicators of the subject's physiological state for the diagnosis of a certain pathology or its evolution analysis [39]. To this end, psychiatrists usually use the Diagnostic and Statistical Manual of Mental Disorders [13], which provides a classification of mental illnesses along with descriptions of the diagnostic categories based on the patient's medical history and the disorders they may show.

Over the past few years, neuroanatomical and neurofunctional analysis have become common practise in the evaluation of certain mental conditions by means of Magnetic Resonance Imaging (MRI), both structural (sMRI) or functional (fMRI), aimed at the study of pathologies and the detection of structural brain anomalies that cause them [11] [51]. For this purpose, different techniques have been proposed in the literature, such as "voxel based morphometry" (VBM) [4], enabling the analysis of structural abnormalities in the brain (see Chapter 2), or the "General Linear Model" [1], which establishes a mathematical model to either analyse sMRI data or obtain the functional response of the brain in fMRI studies.

These research lines have laid the basis for the re-evaluation of previous neuroanatomical hypotheses that were considered to be associated with certain disorders and the proposal of new models with a sound biological foundation. However, in some occasions these results have not been correctly translated to the clinical practise [39]. As a result, there has been a growing interest in the application of other analysis strategies, such as *machine learning* methods, since they are able to describe differences between patient and control groups and to obtain mathematical models that allow discerning between them [33]. The possibility exists that these methods might lead to the establishment of a diagnosis paradigm similar to the score-based systems described above. This would be highly desirable for the psychiatric community.

Machine learning techniques have positioned themselves as some of the most promising options to extract relevant information from neuroimaging data through statistical learning methods. These approaches are mainly characterised by being able to automatically learn a model of data from a collection of examples, which in many occasions can enable the detection of information and data relationships that would otherwise be hidden from the eyes of an expert. For this reason, machine learning methods are being successfully used in data based diagnosis in many fields of medicine. For instance, they are being used in the classification of tissue-cells, the segmentation of retinopathy, the detection of breast-cancer or auricular arrhythmia, just to name a few.

Furthermore, the multivariate nature of these techniques, as well as their ability to extract the greatest amount of available information possible when the number of data is limited, has favoured the widespread use of machine learning tools in neuroimaging analysis [41] and the diagnosis from the information provided by this type of data [29]. This is particularly relevant to our study since our data is composed of a limited amount of sMRI brain scans. So far, scientific production in relation to neuroimaging and machine learning methods has followed a path in which the psychiatric community provides MRI data from an experiment designed to study the brain, and the machine learning community directly applies standard techniques. Because of this, we can find many examples of the application of machine learning approaches to magnetic resonance experiments, such as brain mapping from fMRI data sequences [54], temporal fMRI series analysis [31] or brain state decoding [25] [35]. Clinical applications can also be found, in which the goal is to detect a particular mental illness, such as Alzheimer's disease [53], schizophrenia [12] or obsessive compulsive disorder [47] [40].

Obssesive-compulsive disroder (OCD) is an anxiety disorder that is characterised by recurring intrusive thoughts that induce uneasiness, fear, apprehension or worry as well as repetitive behaviours that are manifested as an attempt to reduce such thoughts. It has significant consequences in the patient's life as the symptoms can be alienating and time-consuming. Its impact can be noticed in the patient's familial, social and professional relationships. It is a chronic psychiatric disorder that affects 2% of the world's population [40].

Prevailing neurobiological models of OCD are based in part on quite solid neuroanatomical findings accumulated over the course of the past years by means of the analysis of structural magnetic resonance imaging (sMRI) data. These findings point to neuroanatomical anomalies that could be associated with the presence and developement of the disease [43]. This makes sMRI brain scans of OCD afflicted patients a perfect candidate for our research.

1.2 Goals and motivations

The vast majority of methods proposed in the literature using machine learning with MRI data focus on analysing differences between patient and control groups. These methods provide a decision on the class to which each MRI belongs in the form of a probability value or a binary value (patient/control), further proving that the images contain relevant information for the diagnosis. In the best cases these studies also provide a subset of voxels or regions that characterize the pathology, which can indicate the psychiatrist or neurologist that a particular region of the brain presents structural or functional differences between healthy and ill subjects. However, given the isolated analysis of these regions in an MRI scan from a single patient, the psychiatrist or neurologist is unable to determine whether the subject is ill or not: the discrimination pattern provided by the classifier comprises, together with these regions and groups, a series of mathematical relations between them that are not directly manageable and are practically impossible to interpret in most cases.

Furthermore, there is an added difficulty in the fact that the available MRI data usually comes from different health centres that employ different magnetic resonance technologies, especially when it comes to the intensity of the magnetic field, producing varying resolution characteristics in the resulting images. On top of this, the image acquisition methods may present differences (for instance, different acquisition sequences like EPI, MEPI or PEPSI) or different space-time bandwidths. For these reasons, studies based merely on voxel or region selection, like VBM, are hard to extrapolate directly from one health centre to another, making it very difficult to find a practical use for them. The cumbersomeness and lack of clarity of the data together with the lack of invariance in the measurement equipment are the main factors that have kept machine learning techniques from being incorporated as practical tools in clinical psychiatry.

The goal of this thesis is to establish a framework for MRI studies using machine learning techniques in such a way that it can eliminate the aforementioned obstacles, making it easier to incorporate machine learning in clinical

psychiatry. To this end we propose a set of models that go beyond the mere classification between healthy and ill subjects and are capable of automatically discovering a set of neural biomarkers, which we will call neuromarkers, that are useful in characterising different mental disorders from the analysis of an MRI brain scan.

A neuromarker is a biomarker that must somehow quantify a neuroanatomical characteristic associated with a pathology. It must also present the following properties:

- Dependence on the subject's endophenotype, such that its values will vary with the pathology subtype and will thus allow the subtype's identification.
- Different values for a given neuromaker must indicate different evolutions of a pathology in different patients, enabling its use in prognosis.
- Patients with different neuromarkers will present varied responses to different medications, making the neuromarkers useful in the prescription process.
- Neuromarkers won't be stationary and will possibly vary with time, which will be an indicator of the patient's evolution.

For these reasons, neuromarkers will be useful in diagnosing, characterising, stratifying, prognosis, prescription and overseeing of a pathology.

With this study we aim to propose a new paradigm to discover neuromarkers capable of characterizing OCD. These biomarkers, named neuromarkers, will be obtained through the automatic analysis of sets of MRI brain scans of OCD patients and control subjects. In order for these neuromarkers to have penetration in clinical psychiatry, they will have to be interpretable and manageable.

The design of these neuromarkers stems from a method for the automatic discovery of clusters of voxels relevant to OCD recently proposed in [40]. With these regions as a starting point, we will first define several candidates to become neuromarkers, that is, we will propose a set of measurements describing features of these individual regions.

In order to obtain a reduced subset of neuromarkers for OCD characterisation, we will apply different selection strategies to remove irrelevant features. This will result in a small set of neuromarkers that is easy to interpret and handle by the psychiatric community.

Experiments will analyse the suitability of each subset of neuromarker candidates, as well as the different selection strategies, showing that we can produce a subset of no more than 50 useful neuromarkers maintaining the original performance in terms of classification error.

1.3 The structure of this thesis

This thesis deals with the application of a series of machine learning techniques on data obtained from structural MRI brain scans. Chapter 2 provides a description of the methods employed to obtain the data. It also introduces the basic notions on machine learning needed to understand the processes that we have designed.

Since our work follows on from the previous research by Doctors Emilio Parrado Hernández, Vanessa Gómez Verdejo and Manel Martínez Ramón, presented in [40], Chapter 3 provides a description of the methods they presented and the implications of their work in the context of our study. Their discoveries and results are the starting point for our research.

Chapter 4 describes in detail the methods designed to obtain a series of useful neuromarkers for OCD characterisation. The first section of this chapter reviews the situation of the problem at this stage. The second section describes the neuromarker types that we have designed. The third section describes the selection methods that we employed to verify the relevance of our neuromarkers.

Chapter 5 presents the results of the experiments. It first analyses the performance of our neuromarkers in classifying healthy people and OCD afflicted patients. It then provides a visual representation of our neuromarkers inside the brain, analysing the significance of the brain regions that have been discovered by providing feedback from the medical community.

The conclusions to this thesis are presented in Chapter 6. Chapter 7

provides a planning and budgetary study for the project.

Figure 1.1 illustrates the whole process, from the initial extraction of the MRI brain scans to the final extraction of relevant neuromarkers and their analysis. It brings context to all the different methods that have been employed throughout the entire research project and can be used as a visual road-map to follow the process as it is described by the text.

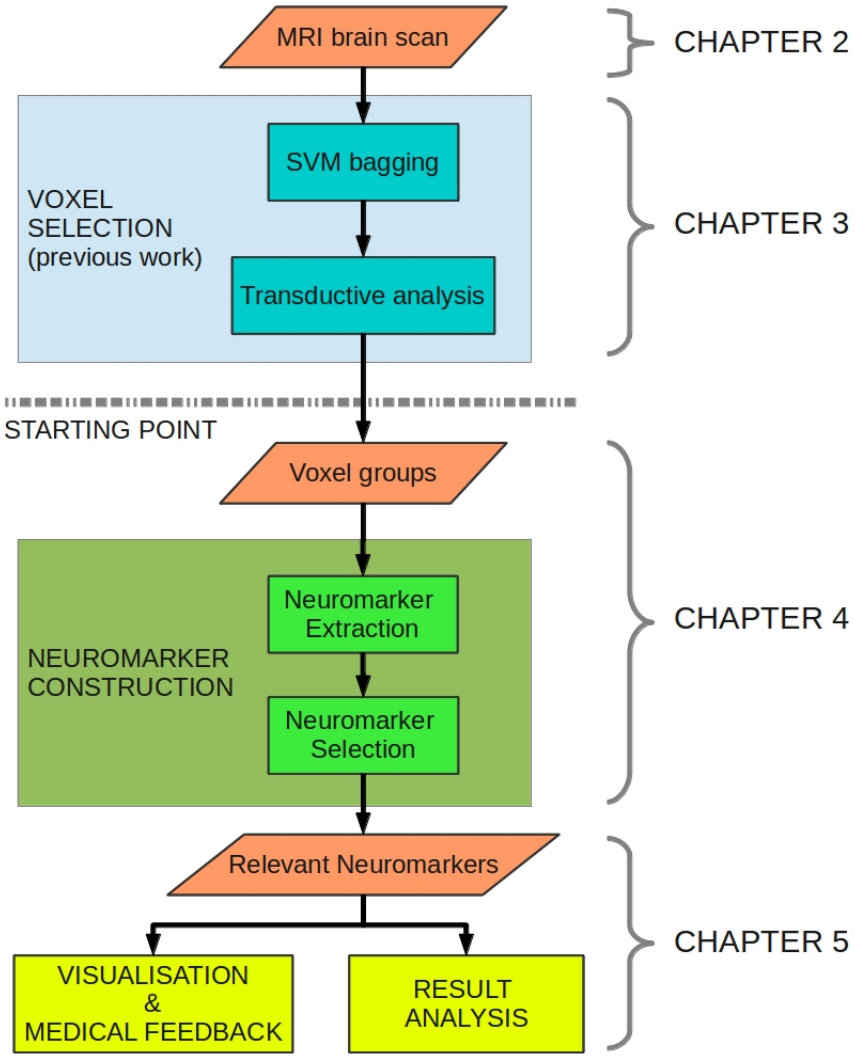

Figure 1.1: Flowchart describing the process described in this thesis.

Chapter 2

The current State of the Art

This chapter introduces some of the most important technologies and techniques that are currently being used in the fields of neural imaging and diagnosis, as well as the machine learning (ML) concepts that are relevant to this thesis.

The first section describes the process of structural magnetic resonance imaging, providing a brief description of how it works and the physical principles it is based on. It then gives a brief review on voxel based morphometry, a popular neuroimaging technique used to locate variations in brain anatomy.

The second section introduces the concept of ML and the basic notions needed to understand the processes and techniques that have been used throughout the project. It gives a few examples of simple ML scenarios and explains the reasoning behind the specific methods and algorithms that are described in Chapters 3, 4 and 5.

The third section describes the support vector machine classifier, one of the most popular ML methods and the one we use in this study.

2.1 Brain structure visualisation

The current diagnosis methods for OCD employ manuals such as the aforementioned Diagnostic and Statistical Manual of Mental Disorders [13]. These manuals establish the procedures to be followed in order to diagnose the dis-

order. Diagnosis based on psychotherapy and pharmacological treatment are the usual lines of action to keep a patient from relapsing. On the other hand, research based on neuroimaging techniques has been made to achieve a deeper understanding of the disorder. The techniques range from structural neuroimaging, such as structural magnetic resonance imaging (sMRI) [5] or computed tomography of the brain (brain CT) [48], to functional neuroimaging, such as functional magnetic resonance imaging (fMRI) [10] and positron emission tomography (PET) [38]. These studies seem to indicate that the origin of the disorder could lie in genetic causes as well as brain anomalies and alterations.

Since the goal of this thesis is to find a series of descriptors of brain regions that are relevant to OCD, we will start with solid neurobiological models that are based on sMRI analysis of healthy and ill subjects.

2.1.1 Structural MRI

SMRI is a relatively new technique that has been used for medical diagnosis since the 1980s [45]. It employs powerful magnetic fields and radio waves so there is no exposure to harmful ionizing radiation forms such as X-rays. This is precisely it's main advantage: it provides a non invasive body imaging method that allows for live analysis. MRI has proven to be efficient in obtaining information on the structure and composition of the body under study. For this reason it is used in a variety of scenarios such as neuroimaging, cardiovascular imaging or musculoskeletal imaging.

The basic operation principle of an sMRI scanner makes use of the magnetic alignment of the hydrogen nuclei (protons) present in the water molecules of the human body. The body is introduced in a strong magnetic field that produces an alignment of the magnetic moments of the protons. A radio field that oscillates at an appropriate frequency is then generated, inducing the emission of a radio-frequency electromagnetic flux by the protons as they go in and out of their magnetic equilibrium state. This flux is then detected by receiver coils that generate a voltage signal that is in turn processed into an image. The orientation of the image can be modified by varying

the main magnetic field. Since the protons in different types of tissue return to their equilibrium state at different rates, structure and composition can be analysed through the intensity differences, or contrast, in the images. Additionally, contrast agents may be introduced intravenously, orally or intra-articularly to further accentuate these structural differences.

A standard sMRI scanner is composed of the following elements:

- A powerful magnet that generates the main magnetic field. Typical clinical-use magnets are super-conducting and cooled by liquid-helium.
- Adjusting "shimming" coils that make sure the magnetic field is stable and homogeneous.
- Gradient coils used to spatially encode the positions of protons by varying the magnetic field across the imaging volume.
- A radio-frequency (RF) system consisting of an emitter subsystem and a receiver subsystem. The emitter is comprised of an RF synthesiser, a power amplifier and a transmission coil. The receiver consists of a receiver coil, a pre-amplifier and a signal processing system. The output of this processing system will be the object of study for this thesis.

The end result is a 3-D image of the brain composed of volumetric pixels or *voxels*. A voxel is simply the elemental volume unit in a three-dimensional image, equivalent to a pixel in a two-dimensional image. In fact, the word voxel is a combination of the words "volume" and "pixel". A voxel contains graphical information associated with a single point in three dimensional space (see Figure 2.1).

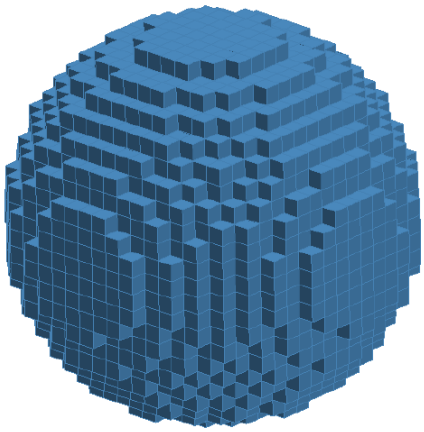

Figure 2.1: Representation of a sphere constructed from volumetric pixels or voxels. Each cube is a voxel containing graphical information.

Each of these voxels represents a numerical value generated by the MRI device's software based on the adjustable input parameters fed to the machine. This information is related to the density of the tissue present in that point in space. We can analyse structural variations in the tissue by observing density variations from voxel to voxel. A typical 3-D representation of a brain obtained with an MRI device can be seen in Figure 2.2 whereas figure 2.3 shows a transversal cut of an MRI brain scan in which three tissue types have been identified from their density.

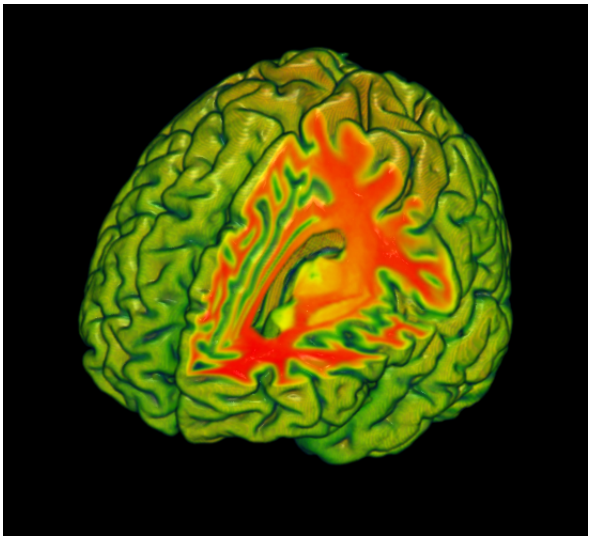

Figure 2.2: Representation of a brain using 3-D graphical information contained in voxels obtained from an MRI device.

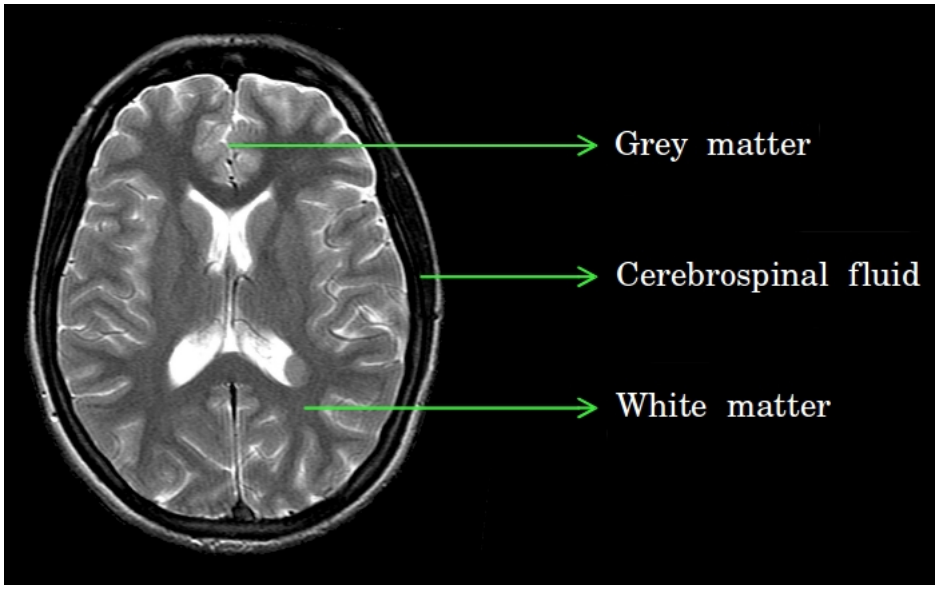

Figure 2.3: Transversal cut of a brain MRI showing segmentation of grey matter, white matter and cerebrospinal fluid.

2.1.2 Voxel-Based Morphometry

Voxel-based morphometry (VBM) [4] is a neuroimaging analysis technique designed to study focal differences in brain structure, using statistical parametric mapping. Traditionally, morphometry measures the volume of the brain or any of its parts by designing regions of interest (ROIs) on brain scans and calculating the enclosed volume. However, this process is complex and time consuming. Furthermore, it is relatively inaccurate, making it unsuitable for measuring the volumes of anything but the largest brain areas, with small volumetric differences going unnoticed.

On the other hand, VBM in it's simplest form involves a voxel-wise comparison of the local concentration of gray matter between groups of subjects. It starts with a spatial normalization of all the brain images in the study into a brain template. It then segments grey matter from the normalised images and applied a smoothing process. Then a series of voxel-wise statistical tests are performed, comparing the grey matter from brains that belong to different groups.

Various VBM studies have been performed searching for different brain functions. In one study, London taxicab drivers were shown to posses a larger than average hippocampus, suggesting a relationship between this particular brain region and spatial awareness and navigation [34]. Another paper [19] used VBM to study the effect of age on grey matter, white matter and cerebrospinal fluid. The study showed that grey matter decreased linearly with age, especially for men, whereas white matter remained roughly constant.

The goal of this thesis is very different from what VBM accomplishes. VBM limits itself to a voxel-wise statistical analysis of structural differences that may be related to a pathology. Our proposition aims at directly relating the presence of OCD to differences in very specific brain regions of patients and healthy people. We also wish to provide a formal representation of these regions so that the difficult and abstract chore of interpreting tens of thousands of voxels is replaced with the analysis of a few, simple neuromarkers that specifically characterise OCD.

2.2 Machine Learning

ML is a branch of the field of artificial intelligence that is concerned with the construction and study of systems that can learn from data. In 1959 Arthur Samuel defined ML as a "field of study that gives computers the ability to learn without being explicitly programmed". In other words, a ML system should be able to perform a certain task without having been told how to do it but rather how to learn to perform the task from experience. This leads us to the more formal definition provided by Tom M. Mitchell: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [37]. This means that a ML system needs to be initially provided with a set of data that will give it experience in performing a desired task. The performance in the realisation of the task will need to somehow be measured in order to tune the learning method and improve the system's capabilities.

Since a ML algorithm depends on the initial training set with which it is provided, and since this set will always be only a sample of the data that is relevant to a task, the algorithm's goal will be to *generalise* as well as possible. Generalisation is defined in this case as the ability of the algorithm to perform accurately on new, previously unseen data that, being of the same type and statistical nature, has not been used in the initial learning or *training* set. In this sense, the measurement of the performance of a learning system must be done on a *test* set that must always be composed of different samples of relevant data from those of the training set.

It follows that the learning process must begin with the system generating a model from the training set using a ML algorithm. It must then measure the accuracy of the model on a test set that is completely separate from the training set.

An example of this process would be a program that is tasked with the prediction of house prices. In a ML scenario an idea for the first step would be to provide the program with a set of houses paired with their known prices. The houses must be defined for the program through one or more relevant

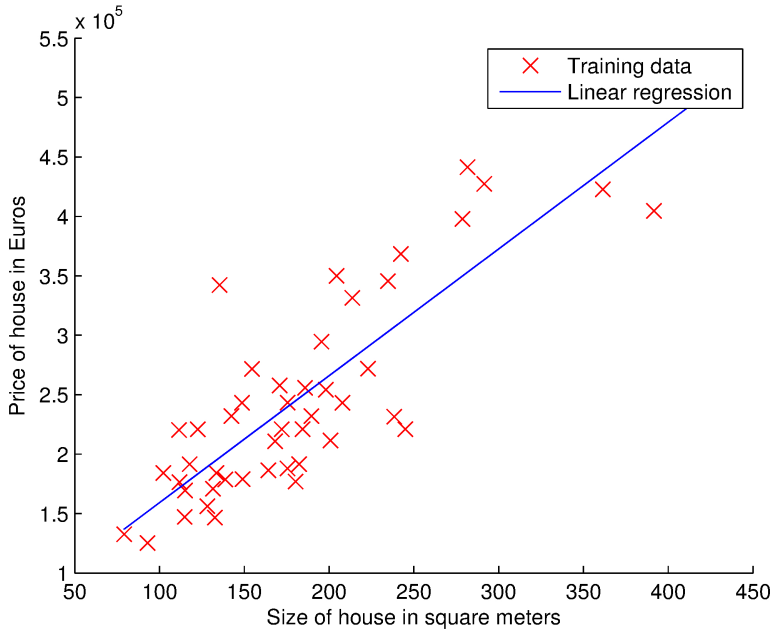

Figure 2.4: Housing prices example. The blue line is a linear model that fits the data.

features. In this example we will use a house's size as the defining feature. The training set would thus be comprised of a number of houses of different sizes for whom the prices are know. The program would then use this set to generate a model that adequately fits the training data. The simplest model in this example would be a linear function of the form $y = w_0 + w_1 x$ where y is the price of the house, x is the house's size and w_0 and w_1 are the parameters of the model used to fit the function to the training data. Next, the houses in the test set would be fed to the model, which would in turn predict a price for each house. It is important to note that the prices of the test houses are not fed to the model, only their sizes. We would then compare the predicted prices with the real ones. The accuracy of the prediction will be our measure of the performance of the system. Figure 2.4 illustrates this example.

2.2.1 Supervised learning and classification problems

The above example illustrates the most simple process that a ML system can follow: it first trains with the training data and then tests its accuracy with the test data. One noteworthy aspect of the example however is the nature of the data that was used: houses were paired with their prices. In other words, the input for the system was a house's size and the output was its predicted price. In order to learn how to perform the prediction, the algorithm needed to train on a training set that provided it with house sizes and their corresponding prices. It then created a model that allowed it to predict the price of a new house given its size. This is an example of what is called *supervised learning*.

In supervised learning each individual subject of data always consists of a pair: the input object comprised of the set of features that defines a subject; and its corresponding label, defined by its desired output. The task is then to predict the label of a new subject given its set of features. In the house pricing example the only feature is a house's size and the label is its price. In this case the labels take continuous values, but there are other cases in which the labels take discrete values. An example of this would be the separation of benign or malignant tumours according to their volume. In this case the task would be to find a classification model that separated benign and malignant tumours as accurately as possible given their volume.

When the supervised learning task deals with discrete labels we call it a *classification* task. The subjects can fall into a series of separate classes and it is the algorithm's duty to define a model that will accurately predict whether a new subject belongs to one class or the other. Tasks that present only two possible classes are called binary classification tasks. There are other situations in which a subject can fall into one of a series of different classes, like for example classifying digital images of galaxies into elliptical, spiral, lenticular or irregular galaxies [16]. Figure 2.5 illustrates a classification example in which tumours are described by both their volume and tissue density and are labelled as "malignant" or "benign". The blue line is a classifier model. New tumours that fall to the left of the line will be classified

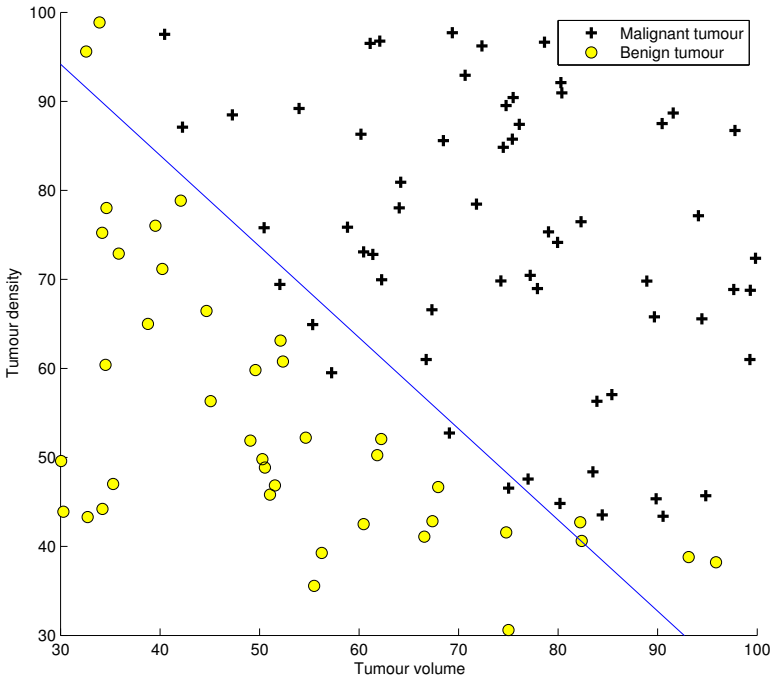

Figure 2.5: Binary classification of benign and malignant tumours according to their volume and tissue density.

as benign, whereas those that fall to the right of the line will be classified as malignant.

There are other cases in which the data is not specifically labelled. All we have is a mass of raw data and we need to find some structure within it that can help us describe it. This is called *unsupervised learning*. Since we don't have a labelling pattern we can't devise an error or reward signal to evaluate a potential solution. What we can do is find relationships between subjects that produce structured groupings or clusters in the data. These techniques are currently being used in many applications such as genome sequencing and gene function definition [28].

In the case of this project we are dealing with a binary classification task: do the MRIs belong to healthy or to OCD afflicted subjects? As will be further described in Chapter 3, we start with a collection of MRI

scans belonging to control subjects which are known to be healthy and OCD patients which have been positively diagnosed with the illness. It is thus a supervised binary classification learning problem.

2.2.2 Avoiding data overfitting

When designing a machine learning system, it is of vital importance that the training data set and the testing data set are composed of different samples of the relevant data type. The reason behind this imposition is that one can always design a model that perfectly fits the data. In the house prices example, one could easily arrive at a model like the one presented in Figure 2.6 by using a non linear function of the variables that is forced to adapt very well to the training samples. If we then use the training data to test our error rate the result will be very satisfactory since our non linear model predicts the price of each house in the training set very well. But is this not a realistic prediction model. The truth is that if we used such a model on new houses to predict their price the results would be very poor. The error rate we obtained is constrained to a very small sample of the enormous amount of possible houses one can find in the real world. Our model is therefore unable to generalise well when it comes to dealing with reality. This is what is known as *data overfitting*.

Generally speaking, whenever we see that a predictor obtains a very low training error rate but performs badly when tested on new data we are very probably overfitting our training data. In some scenarios overfitting is very difficult to avoid. Specifically, whenever a problem presents a very small number of available training samples with a very large number of defining features, there will be a very high risk of overfitting. This is known as the *small sample problem* and it is very important in the case of this study. The methods described in Chapter 3 deal with strategies to alleviate its impact.

Another instance in which overfitting can appear is when a specific learning strategy presents a control parameter (or set of control parameters) that can be varied to modify its behaviour and performance. These parameters can range from modifiable constants inherent to the mathematical formula-

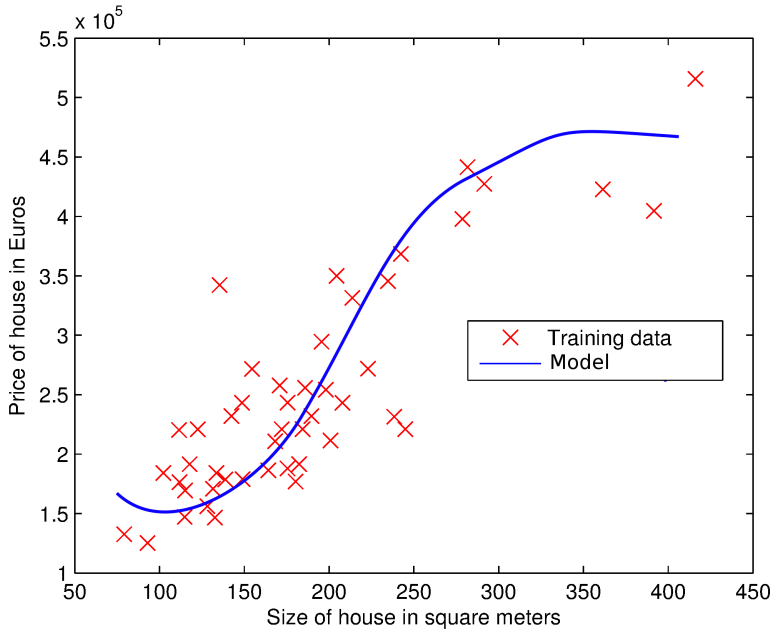

Figure 2.6: Housing prices example. The blue line is a non-linear model that overfits the training data.

tion of the ML method, to the number of variables we use in training, or the number of performed iterations of an algorithm.

Any situation in which a control parameter is modified and tested is prone to produce overfitting unless we are very careful with what data we use to perform the tests. If we test the parameter against the training data we will be optimizing it so that it performs well with the training set but, again, it will not necessarily generalise well to the test set. In this case a *validation* process must be performed. To validate a parameter we must obtain a separate validation set by, for example, obtaining new data, splitting the original training set into a smaller training set and a validation set or employing other validation strategies such as cross-validation, K-fold validation, etc. The precise validation process that was employed during the development of this thesis is described in Chapter 5.

2.2.3 Feature selection

When attempting to solve a classification problem by means of machine learning methods, a common obstacle that researchers encounter is the large number of input variables that the data presents. In the field of MRI studies, the number of available subjects is generally very small compared to the number of variables, in the form of voxels. Chapter 3 explains how this issue is dealt with in the particular case of this study.

But having too many variables presents difficulties beyond the small sample problem, depending on the nature of the case: it often complicates the visualization and understanding of the data, it increases measurement and storage requirements and it increases training and utilisation times. Feature selection strategies aim at finding the subset of features that is most relevant and informative. In many cases, a data set will present redundant or irrelevant features that do not provide information useful towards classification and are therefore no more than noise [23].

The most often used feature selection strategies fall into three main categories, presented here in order of complexity [22]:

- **Filters:** Filters use relevance measurements to analyse how useful each individual feature is. The selection of features is thus independent from the classification task and happens as a preprocessing stage prior to the training stage. Relevance criteria can be combined with search algorithms to produce subsets of variables. The criteria can also be directly used to produce a ranked list of variables. Two of the most commonly used search algorithms are *forward search* and *backward search*. Backward search algorithms start by evaluating the relevance criterion with all the features, and then proceed to eliminate the least relevant features one at a time in decreasing order of relevant. Forward search algorithms on the other hand start with a single feature, the one found to be individually most relevant, and add one feature at a time on each iteration according to its relevance in decreasing order. Both forward and backward search algorithms provided *nested subsets* of variables, with each subset providing higher relevance than the next. Chapter 4

Figure 2.7: Filter selection method.

presents specific descriptions of how these two search algorithms were used in our study. Figure 2.7 illustrates the basic idea behind a filter. Filters are computationally light and therefore tend to be fast.

- **Wrappers:** Wrappers use ML to select relevant feature subsets. They train a classification model with different feature subsets and produce a ranking based on the classification error obtained with each one. The fact that they work alongside the classifier allows them to obtain feedback from the classification output. A wrapper needs to define a search algorithm. While exhaustive search methods can be used with small data sets, the problem quickly becomes computationally very intensive [2]. Also, exhaustive selection of variables is prone to produce overfitting of the training data since the method tends to pick those features that produce good results with a given training set, but not necessarily in a more general scenario. This can again be alleviated by the use of forward or backward search algorithms, as well as other search methods. In the case of our study, the use of a wrapper method even in combination with a backward search algorithm proved to be too computationally intensive. Figure 2.8 depicts a wrapper method that implements a backward search algorithm.

- **Embedded methods:** Embedded methods are algorithms that are directly integrated into the classification task. Feature selection is applied alongside the classification task, selecting those features that seem to improve performance. Since they are embedded in the classifier, their nature depends on the specific classification method used. The recursive feature elimination algorithm described in Chapter 4 is an example of an embedded selection method. These methods tend to fall in between filters and wrappers as far as computational intensity

is concerned. They are also less prone to overfitting than wrapper methods.

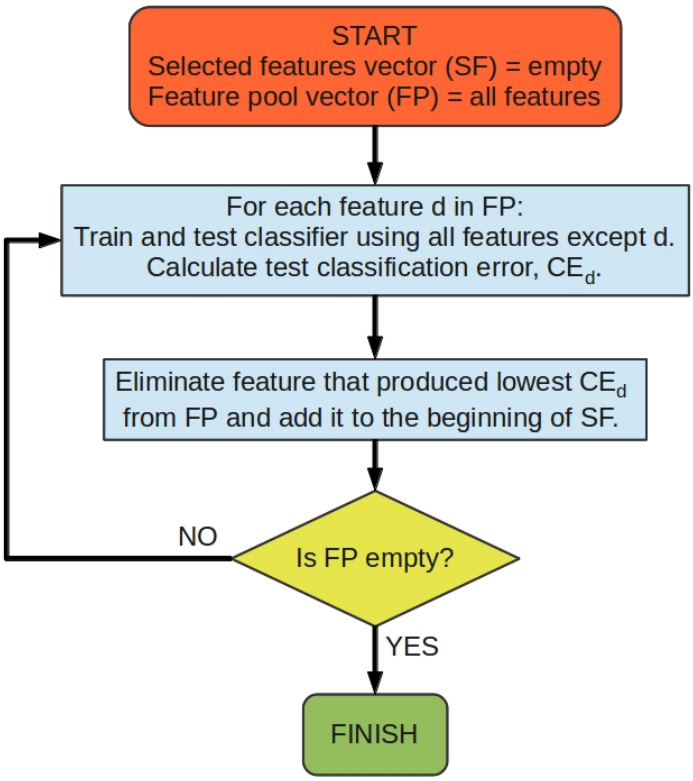

Figure 2.8: Wrapper selection method.

As was stated in Chapter 1, the goal of this thesis is to offer a collection of biomarkers that is easy to use and interpret. To this end, various filters and one embedded method have been used in the selection of relevant neuromarkers once these have been defined (see Chapters 4 and 5).

2.3 The Support Vector Machine classifier

Classification models divide the feature space into disjoint regions assigned to class labels [23]. To this end numerous classifying models have been devised and applied with success over the years. Good examples are logistic regression [26], neural networks [7], the K nearest neighbours algorithm [15],

decision trees [42], random forests [9] or gradient boosting algorithms [17]. However, out of all the available classification strategies, perhaps the most widely used, and the initial go-to choice, when one is presented with a difficult classification problem is the *support vector machine* (SVM) classifier [49].

The basic mathematical idea for SVMs was initially presented in the Generalized Portrait algorithm by V. Vapnik and A. Lerner in the nineteen-sixties while they were working at AT&T Laboratories. In 1979, Vapnik wrote the book *Estimation of dependences based on empirical data*, translated to english in 1982 [52]. In this book, apart from setting the foundation for the statistical theory of learning and generalisation, he introduced a generalisation of the Generalised Portrait algorithm that would end up being developed into the SVM classifier. In 1992 B. Boser, I. Guyon and Vapnik published *A training algorithm for optimal margin classifiers* [8] in which a formal definition of the SVM was established. Later developments like C. Cortes' and Vapnik's *Support-vector networks* [14] propelled the SVM to its current popularity by providing at least an equal level of performance to other state of the art techniques such as neural networks.

2.3.1 The linearly separable case

Initially, SVMs were conceived to classify data belonging to two classes that were linearly separable. This means that a "gap" exists between the data of the two classes such that they can be perfectly separable by a single hyperplane of the form:

$$\mathbf{w} \cdot \mathbf{x} - b = 0, \tag{2.1}$$

where $\mathbf{w}$ is the weight vector normal to the hyperplane and b is the *bias term* where $\frac{b}{||\mathbf{w}||}$ determines the offset of the hyperplane from the origin along $\mathbf{w}$. It can be easily seen that in the case of a linearly separable problem there is an infinite number of hyperplanes that satisfy the classification requirements (see Figure 2.9).

In the case of an SVM, we are looking to maximize the distance between the hyperplane and the nearest point of any class. This distance is called the

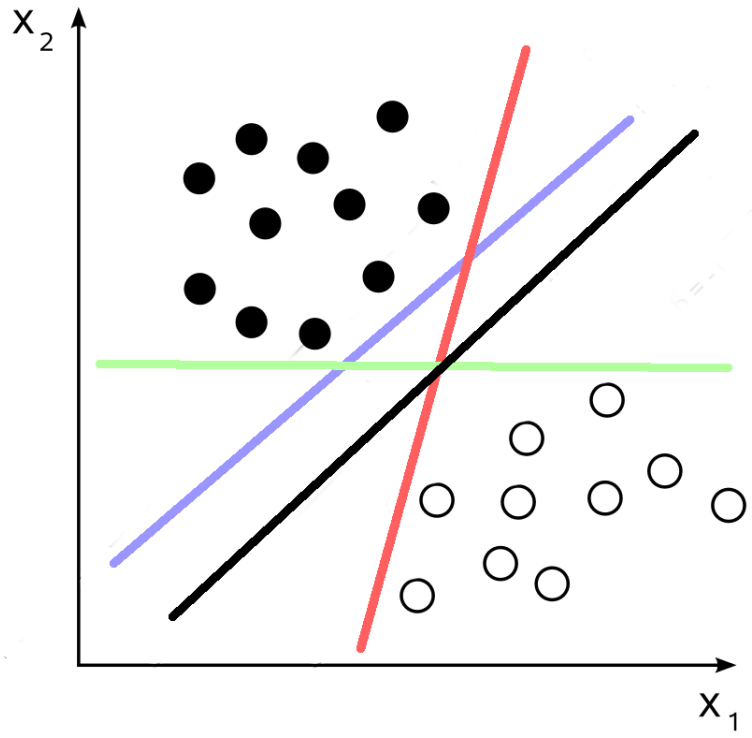

Figure 2.9: A linearly separable two dimensional case: the coloured lines represent some of the infinite hyperplanes that correctly separate the data.

functional margin. It can be generally stated that the larger this margin is, the better the classifier will generalise. This margin can be found by selecting two hyperplanes that separate the data and then maximizing the distance (the margin) between them. The two hyperplanes are defined by

$$\mathbf{w} \cdot \mathbf{x} - b = 1 \tag{2.2}$$

and

$$\mathbf{w} \cdot \mathbf{x} - b = -1. \tag{2.3}$$

The margin between the two hyperplanes is geometrically defined as $\frac{2}{||\mathbf{w}||}$. It follows that in order to maximize this margin we need to minimize $||\mathbf{w}||$.

Equations (2.2) and (2.3) can be rewritten to prevent data points from

falling into the margin as:

$$\mathbf{w} \cdot \mathbf{x}_m - b \geq 1, \text{ when } y_m = 1 \tag{2.4}$$

and

$$\mathbf{w} \cdot \mathbf{x}_m - b \leq -1, \text{ when } y_m = -1. \tag{2.5}$$

where $\mathbf{x}_m$ is a subject in the training set and y_m is its corresponding label.

In this case, the classification problem is to minimise $||\mathbf{w}||$.

The optimisation problem is simplified by substituting $||\mathbf{w}||$ with $\frac{1}{2}||\mathbf{w}||^2$. This substitution leaves the solution intact and eliminates the square root operation implicit in $||\mathbf{w}||$. The constrained optimisation problem is thus formulated by:

$$\arg\min_{(\mathbf{w},b)} \tfrac{1}{2}||\mathbf{w}||^2 \tag{2.6}$$

$$\text{st.} \quad y_m(\mathbf{w} \cdot \mathbf{x}_m - b) \geq 1 \quad \text{for } m = 1 \ldots M$$

where the constriction is the unified expression of (2.4) and (2.5).

2.3.2 The non-linearly separable case: soft margin SVMs

Since there are many classification problems where the training data is not linearly separable (the distributions of both classes overlap), the SVM needs to be modified. The solution in this case implies finding a hyperplane that makes as few classification mistakes as possible while still maximizing the margin between the hyperplane and the nearest cleanly classified data. This *soft margin* method introduces non-negative slack variables, ξ_m, which measure the degree of misclassification or error penalty for each $\mathbf{x}_m$.

In this case, the equivalent to (2.6) with introduction of the error variables is:

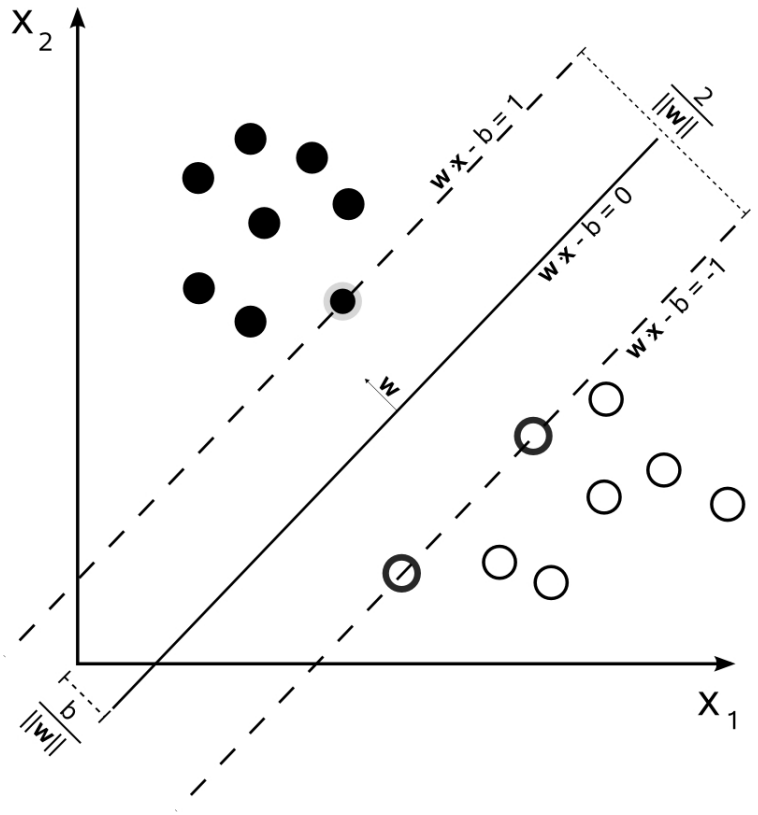

Figure 2.10: Classification margin of an SVM in a linearly separable case.

$$\arg\min_{(\mathbf{w},\xi,b)} \left\{ \tfrac{1}{2}||\mathbf{w}||^2 + \mathrm{C} \sum_{m=1}^{M} \xi_m \right\} \tag{2.7}$$

$$\begin{aligned} \text{st.} \qquad & y_m(\mathbf{w} \cdot \mathbf{x}_m - b) \geq 1 - \xi_m \qquad && \text{for } m = 1 \ldots M \\ & \xi_m \geq 0 && \text{for } m = 1 \ldots M \end{aligned}$$

Non zero ξ_m are penalised and the optimisation problem is now a trade-off between maximizing the margin and minimizing the error penalty. C is a parameter that regulates the trade-off between margin maximisation and error minimisation.

Both the linearly separable and the non-linearly separable cases can be solved by using Lagrange multipliers. In the case of the non-linearly separable

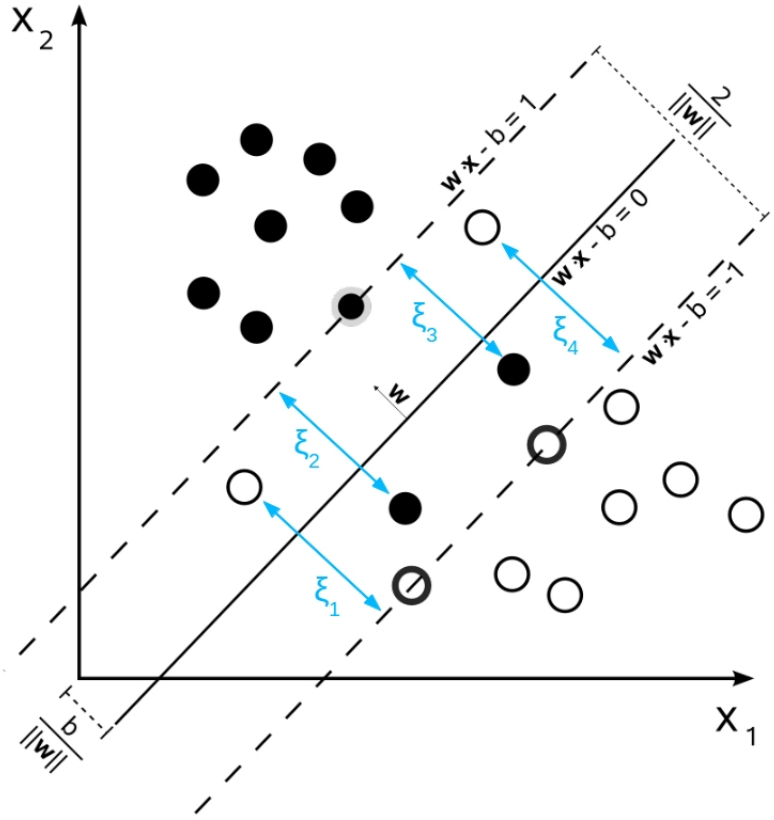

Figure 2.11: A non-linearly separable case solved using a soft margin SVM.

problem, the formulation using the Lagrange multipliers α_m and β_m is as follows:

$$\arg \min_{(\mathbf{w},\xi,b)} \max_{(\alpha,\beta)} \left\{ \frac{1}{2}||\mathbf{w}||^2 + \mathcal{C} \sum_{m=1}^{M} \xi_m - \sum_{m=1}^{M} \alpha_m [y_m(\mathbf{w} \cdot \mathbf{x}_m - b) - 1 + \xi_m] - \sum_{m=1}^{M} \beta_m \xi_m \right\} \tag{2.8}$$

where the goal is to find the saddle point that minimizes $\frac{1}{2}||\mathbf{w}||^2$ and ξ_m and maximizes α_m and β_m, for $m = 1 \dots M$. It is important to note that all the points for which $y_m(\mathbf{w} \cdot \mathbf{x}_m - b) - 1 > 0$ do not affect the solution since their corresponding α_m will be set to zero.

The solution can now be expressed as a linear combination of the training vectors:

$$\mathbf{w} = \sum_{m=1}^{M} \alpha_m y_m \mathbf{x}_m, \tag{2.9}$$

where only a few of the α_m will be greater than zero. These correspond to the samples that satisfy $y_m(\mathbf{w} \cdot \mathbf{x}_m - b) = 1$, meaning that they reside exactly on the maximum margin. These samples are called the *support vectors*. In Figures 2.10 and 2.11 we can see that the two hyperplanes $\mathbf{w} \cdot \mathbf{x} - b = 1$ and $\mathbf{w} \cdot \mathbf{x} - b = -1$ rest on just a few of the training samples: the support vectors. The classifier hyperplane, $\mathbf{w} \cdot \mathbf{x} - b = 0$, lies exactly in between.

The binary classification estimation function for a new subject $\mathbf{x}$ can now be defined as follows:

$$\hat{y}(\mathbf{x}) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m y_m (\mathbf{x}_m \mathbf{x}) + b\right) \tag{2.10}$$

Chapter 3

Previous work: a review on discovering brain regions

This chapter is a review on the work of doctors Emilio Parrado Hernández, Vanessa Gómez Verdejo and Manel Martínez Ramón, in collaboration with medical personnel of the Bellvitge University Hospital, Barcelona (Spain). Specifically, it goes over the key aspects of the paper "Discovering brain regions relevant to obsessive-compulsive disorder identification through bagging and transduction" [40], published in 2014 in the Medical Image Analysis journal.

This bachelor thesis stems from this work and is a continuation of the efforts made by the research team to find a solution to OCD characterisation through ML techniques.

3.1 Initial data description and preprocessing

Eighty-six outpatients with OCD (44 males; mean ± SD age, 34.23 ± 9.25 years) were recruited from the outpatient service of the Department of Psychiatry of the Bellvitge University Hospital. Diagnosis was confirmed by two senior psychiatrists through separate interviews, held one month apart, using the Structured Clinical Interview for DSM-IV Axis I Disorders (First et al., 1997). All the patients had experienced OCD symptoms for at least one year

prior to the assessment and none of the patients met criteria for Tourette syndrome, psychotic disorder or psychoactive drug abuse/dependence. The presence or past history of any neurological or serious medical condition, in addition to the presence of any sign of abnormality in the MRI scan, were also regarded as exclusion criteria. Comorbid major depression and anxiety disorders were not considered to be exclusion criteria provided that OCD was the primary diagnosis.

A group of 86 healthy control subjects from the same sociodemographical environment was also recruited. Control subjects were selected according to the same exclusion criteria and did not differ from the patient group in age or gender distribution (43 males; mean $\pm$ SD age, 33.47 ± 9.94 years).

The brain images were acquired with a 1.5 Tesla Sigma Excite system (General Electric, Milwaukee, Wsiconsin) equipped with an eight channel phased-array head coil. A high T1-weighted anatomical image was obtained for each subject using a 3-dimensional, fast spoiled gradient inversion-recovery prepared sequence with 130 contiguous slices (TR = 11.8 ms; TE = 4.2 ms; flip angle = 15^{o}; field of view = 30 cm; 256×256 pixel matrix; slice thickness = 1.2 mm).

The notation that will be followed from now on throughout this thesis (unless stated otherwise) is defined as follows:

- $\mathbf{X}$ is the data matrix of size $M \times D$ containing all subjects.
- M is the total number of subjects in our data, both patients and controls.
- D is the total number of voxels per subject.
- Each row of $\mathbf{X}$ is a vector, $\mathbf{x}_m$, of size D that describes subject m through its D voxels.
- Vector $\mathbf{y}$ is the label vector containing the labels for all the subjects. Control subjects are labelled as $y = -1$ while patients are labelled as $y = 1$.

An important consideration needs to be made when working with sets of sMRI brain scans: the analysis of each individual image is performed under the assumption that its voxels are localised in the same anatomical regions as in the rest of the images. Since no two brains are alike, a certain standardisation of the brain scans needs to be performed before this assumption can be made. Thus, several pre-processing stages need to be applied to each and every image. These techniques are commonplace in MRI analysis processes such as VBM, and usually include a tissue segmentation stage, an image normalisation stage to a common anatomical template and a final image smoothing stage. The exact processes that were applied in this case are:

- **Segmentation:** The first stage is to segment the image, dividing the brain into the various types of tissue it's composed of. This transformation simplifies image analysis and interpretation. Segmentation is achieved by estimating the probability that each voxel contains matter of one tissue type or another. Since each voxel represents the density of the tissue it contains, and the densities of each tissue type follow known Gaussian probability distributions, the probability that a voxel contains one type of tissue or another can be obtained by comparing its value to the different tissue density distributions. In our case we are interested in the probability that a voxel contains grey matter.

- **Normalisation:** The second and most important stage is a normalisation of all the brain images. This is key, since all the experiment's data must belong to the same variable space. Normalisation must adjust the anatomy of each subject's brain to a standardised brain template, like the "MNI brain", without incurring in any significant distortion. In this case, the data was processed using the DARTEL approach, which applies a series of non linear transformations that adapt each image to a common framework [3]. Additionally, the Jacobian determinants derived from this spatial transformation were used to modulate the image voxel values and restore the volumetric information lost during the normalisation process.

- **Spatial smoothing:** Spatial smoothing is the final stage of pre-processing.

In this stage the image is filtered using a 4 mm full-width at half-maximum (FWHM) Gaussian kernel. This process is useful since it cleans the image of noise and simplifies statistical analysis.

3.2 Finding relevant voxels

The above preprocessing method provides a series of grey matter segments producing a vector of positive coordinates associated with the probability that the corresponding voxel is composed of grey matter. Due to the fact that in our problem the number of input variables (of the order of 10^5 voxels) is far greater than the number of subjects (a total of 172) independently of the label assigned to each subject (see Section 2.2 for a description of the *small sample problem*), statistical learning theory principles dictate that the problem is fully separable by a linear classifier.

A linear classifier implements a discriminant function of the following form:

$$\hat{y}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b) = \text{sign}\left(\sum_{d=1}^{D} w_d x_d + b\right), \tag{3.1}$$

where $\mathbf{w} = [w_1, \ldots, w_D]^T$ and b are the weight vector and the bias term, respectively.

Since voxel grey-matter probability is always positive, a linear classifier allows for a straight forward interpretation of the role that each voxel plays in the classification process by analysing the value of its corresponding element in the weight vector. Each voxel can be classified into one of the following groups in accordance with its subject's label:

1. A voxel whose w_d takes a relatively high negative value will likely be indicative that $\mathbf{x}$ is a healthy subject.

2. A voxel whose w_d takes a relatively high positive value will likely be indicative that $\mathbf{x}$ is an OCD patient.

3. A voxel whose w_d takes a negligible or zero value indicates that it is not relevant to estimating the correct label of $\mathbf{x}$.

It follows that a simple voxel selection process would keep the voxels that fall under the first two categories and discard the voxels that fall under the third category.

3.2.1 Bagged Support Vector Machines for voxel selection

The first natural choice for a linear classifier is a soft-margin linear SVM. Recalling what was stated in Section 2.2, in a soft -margin SVM $\mathbf{w}$ and p are calculated by solving the optimization problem:

$$\arg\min_{(\mathbf{w},\xi,b)} \left\{ \tfrac{1}{2}||\mathbf{w}||^2 + \mathrm{C} \sum_{m=1}^{M} \xi_m \right\} \tag{3.2}$$

$$\begin{array}{lcl} \text{st.} & y_m(\mathbf{w} \cdot \mathbf{x}_m - b) \geq 1 - \xi_m & \text{for } m = 1 \ldots M \\ \text{st.} & \xi_m \geq 0 & \text{for } m = 1 \ldots M \end{array}$$

The starplots method proposed in [6] uses an interesting approach in order to achieve good generalisation in problems with sparse training data. In this method, an ensemble of diverse linear classifiers is constructed. Then the voxels that present consistent weight patterns over the classifier pool can be determined as relevant to the classification task. This ensemble is constructed on the basis of *bagging classifiers*. The bagging classifier approach picks a subset of L subjects from the training data set at random and without replacement. It then trains the classifier and calculates the weight vector, $\mathbf{w}$. It repeats this process S times where S is the number of classifiers in the ensemble, each time storing the classifier's weight vector. It then checks the sign-consistency of the voxels over the classifiers in the ensemble and groups them into the following categories:

1. The voxels with w_d taking a positive value in $r\%$ of the classifiers.
2. The voxels with w_d taking a negative value in $r\%$ of the classifiers.
3. The voxels that do not show the sign-consistency needed to fall into either of the other two groups.

The consistency threshold r must be properly validated.

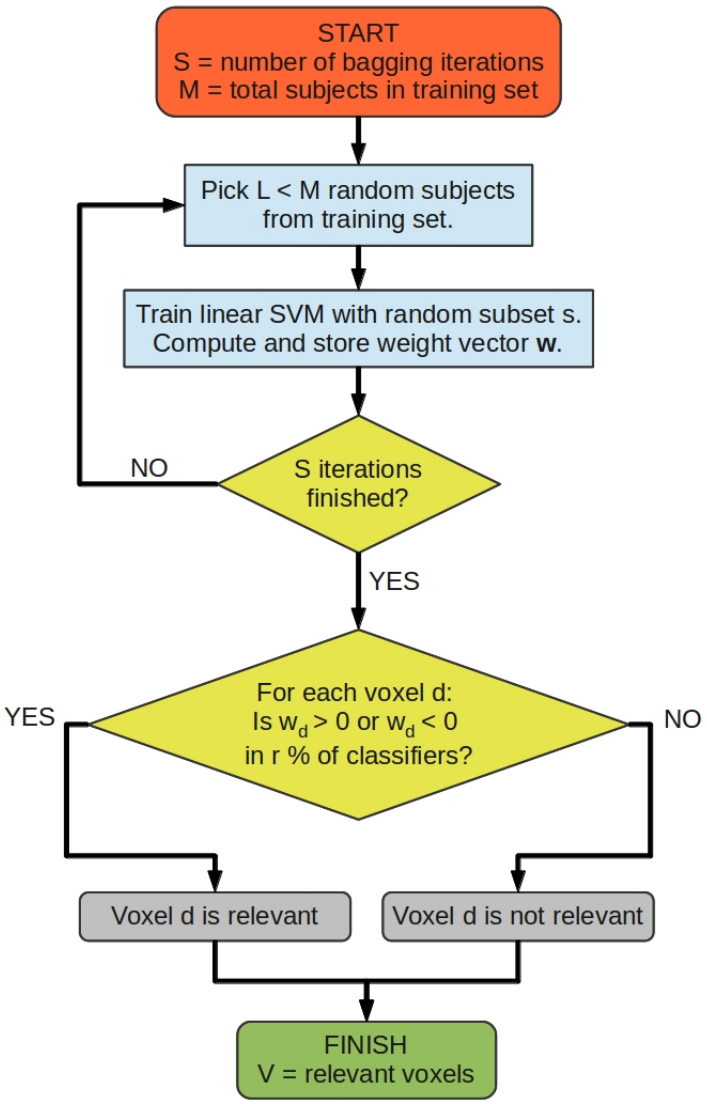

Figure 3.1: Flow chart depicting the SVM bagging process. M is the total number of subjects. L is the number of subjects in each classifier. S is the size of the classifier ensemble. V is the subset of voxels that have been found to be relevant.

Once the bagging process concludes, only the voxels that fall into the first two categories are considered to be relevant to OCD characterisation since they consistently show up as aligned with the classification process. Figure

3.1 illustrates the SVM Bagging process.

3.2.2 Transductive refinement of the voxel selection

Due to the extreme nature of the small sample problem presented by our data, it is very difficult to determine which of those voxels selected by the bagging process are truly related to OCD and which are being selected simply because they are capable of separating a particular brain from the others, regardless of whether its label is positive or negative.

In order to deal with this problem the techniques of *transductive learning* and *conformal analysis* offer a series of ideas and strategies that can prove to be very helpful towards variable selection. The basic idea behind transductive learning is that forcing labels of test subjects to be positive or negative (regardless of what their original label was) and then training the classifier can provide us with information as to whether certain variables consistently help the classifier perform its task or, by contrast, they are highly dependant on the initial labelling of the subjects.

Applying this idea to OCD characterisation, the team developed the following refinement to the SVM bagging strategy. The process starts with the selection of a single test subject from the data set, $\mathbf{x}_{test}$. The SVM bagging process will then be performed twice, with the label for $\mathbf{x}_{test}$ being forced to be positive in the first iteration and negative in the second. This will give us two sets of relevant voxels:

- V_+ is the set obtained in the bagging process that had $\mathbf{x}_{test}$ labelled as positive.
- V_- is the set obtained in the bagging process that had $\mathbf{x}_{test}$ labelled as negative.

The final selection of voxels will be the intersection between V_+ and V_-. The intuition behind this reasoning is that voxels that appear in one of the two subsets, but not in the intersection, are highly dependant on the particular labelling of $\mathbf{x}_{test}$ but are not necessarily relevant to OCD. This could be due to factors that have nothing to do with the disease such as

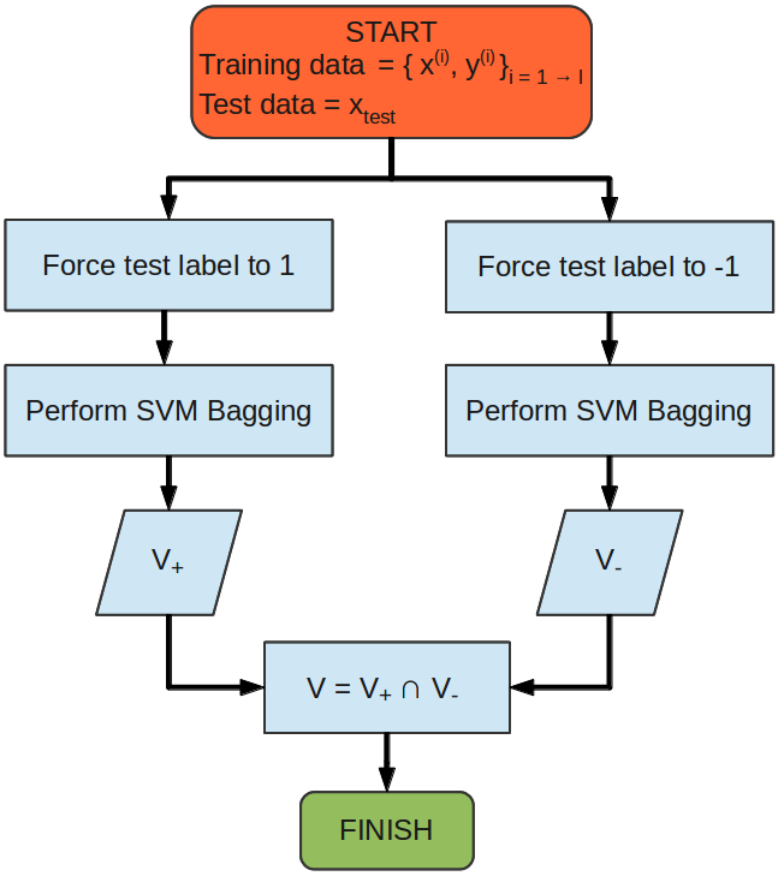

Figure 3.2: Flow chart depicting the T-BS process.

gender, age, etc. Figure 3.2 illustrates the transductive refinement of SVM bagging (T-BS).

This process is repeated once for each subject in the test set. This provides a different subset of voxels for each test subject. A classifier is then trained and tested with each of the subsets and the test errors are averaged, resulting in a global classification error.

3.3 Initial results and conclusions

After the experiments were performed, the empirical results showed that the T-BS process is very effective at selecting a set of approximately 43.000

voxels that are strongly related to OCD. However, this set of voxels is by itself not particularly helpful in clinical applications since clinicians characterize diseases in terms of region-wise features. For this reason a clustering process of interconected voxels was applied, resulting in an average of 718 ± 40 groups.

In order to check whether or not the detected regions were reliable, the classification accuracy of the T-BS approach was compared to the accuracy of two other base-line variable selection approaches: mass-univariate voxel selection and recursive feature elimination [23]. The classification error of these methods ranges between 32% and 37%. The advantages of the T-BS method became obvious when it was observed that it obtained a classification error of 26.2%.

To summarize this chapter, the T-BS method applied to voxel selection led to the identification of a large set of OCD related brain alterations. After the connected voxels are adequately clustered, these alterations can be characterised in terms of region wise features of clinical relevance. Furthermore, the transductive refinement greatly improved the control of the small sample problem. However, the classification accuracy is still too low for clinical purposes. Also, the resulting data is still too abstract in nature to be easily managed by the psychiatric community. Some transformation of the data needs to be performed in order to achieve a more meaningful interpretation of these 43.000 voxels. The following chapters will introduce and experiment with various ideas aimed at constructing a set of biomarkers and thus achieving this goal.

Chapter 4

Building and selecting Neuromarkers

This chapter presents a formal description of the bulk of the work that has been performed during the development of this thesis.

It begins with a revision of the relationship between the investigation and its medical usefulness at this point. It illustrates the necessity to process the data into a reduced and useful set of features in the form of neuromarkers. It then proceeds with a definition of the different ideas that we have implemented in order to find these neuromarkers. Finally, it deals with the description of the feature selection methods that have been developed to further reduce the set of relevant neuromarkers that is most clinically-friendly and interpretable by the psychiatric community.

4.1 Motivations and goals

There is no doubt that the process described in the previous chapter is invaluably useful. It must be noted that the initial dimension of the data set was of the order of 500.000 variables in the form of voxels. This meant that the dimensionality of the data was three orders of magnitude greater than the number of subjects we had to work with. After the voxel selection process, we are left with around 43.000 voxels that are relevant to OCD characteri-

sation. Not only is this reduction of one order of magnitude substantial, but the classification error also rate dropped from around 40% when using all the voxels to just over 26% after the T-BS process. However, despite the usefulness of the aforementioned process, the obtained subset of voxels presents an unfriendly characterisation of the OCD pathology, since its huge size is still unmanageable and difficult to understand for the clinical community. It is almost impossible to relate the value of each individual voxel with the brain deformity or dystrophy which may be related to the disorder.

However, we can exploit the grouped distribution of these voxels, the 718 brain regions on average described at the end of Chapter 3, to define a set of measurements which are able to represent the relevant information of these brain regions in a friendly manner. Due to the fact that these measurements must be useful for disease characterisation, we will denote them as neural biomarkers, or simply neuromarkers, as per the definition of a biomarker presented in Chapter 1. Section 4.2 of this chapter describes the methods that have been applied to the data in pursuit of valuable neuromarkers.

Once the neuromarkers have been constructed, the goal will be to verify whether they are all relevant to the task of OCD characterisation or whether some of them can be ruled out. This would reduce the number of features even more, simplifying the problem and leaving us with a much more manageable set of neuromarkers. In order to measure the relevance of our neuromarkers we have implemented a series of feature selection algorithms, described in Section 4.3.

The verification of the validity of these processes will be performed through validation, as will be described in Chapter 5. It is important to stress once again that our goal now is not necessarily to improve the classification error, but rather to arrive at a small set of useful features that adequately characterise an OCD afflicted brain. We even consider that a small increase in the classification error is an acceptable trade-off if it implies an important reduction of features that makes the data more intuitive and easy to interpret.

4.2 Building Neuromarkers

First we will devise a series of methods to define different types of neuromarkers. Each voxel of an MRI scan is characterized by its grey matter probability. We have grouped the relevant voxels in a subset of 718 brain regions on average over all the subjects, with each region noted by S_g, $g = 1, \ldots, G$. We can now characterize the grey matter probability of these regions with some measurements or data transformations that we will define as neuromarkers.

After a series of trials, we have arrived at four different types of measurements that can be used as neromarkers for OCD characterisation, providing a single parameter for each brain region that represents the entirety of the voxels it comprises. The methods used to construct these neuromarkers are described in the following subsections.

The validity of each of these four neuromarkers will be tested in Chapter 5. It is important to note that each of these four types of neuromarker acts independently from the others and that the task will be to discern which one is better suited to the task of OCD characterisation.

4.2.1 Average of grey matter probability

This first, most basic, measurement directly obtains a single parameter for subject $\mathbf{x}_m$ over each brain region S_g by averaging the gray matter probability values of the voxels that belong to it. We will refer to this neuromarker as the AV neuromarker from now on:

$$\mathrm{AV}_m^{(g)} = \frac{1}{|S_g|} \sum_{i \in S_g} x_{m,i} \tag{4.1}$$

where $\mathrm{AV}_m^{(g)}$ is the AV neuromarker for region S_g and subject $\mathbf{x}_m$, and $|S_g|$ is the number of voxels in brain region g.

This neuromarker can be understood as a representation of the surface area of each brain region.

4.2.2 Accumulated grey matter probability

Another interesting parameter can be obtained by summing the values of all the voxels of subject $\mathbf{x}_m$ belonging to each region S_g. We will refer to this neuromarker as the ACC neuromarker from now on, given by:

$$\mathrm{ACC}_m^{(g)} = \sum_{i \in S_g} x_{m,i} \tag{4.2}$$

where $\mathrm{ACC}_m^{(g)}$ is the ACC neuromarker for region S_g and subject $\mathbf{x}_m$, and $|S_g|$ is the number of voxels in brain region g.

Note that, unlike AV markers, this marker is not dividing by the brain region size. Therefore, it can be interpreted as a representation of the volume of each brain region.

4.2.3 Variance of grey matter probability

Here, we consider the variance of the voxel gray matter probability to represent each brain region. The idea of this neuromarker is that evaluating variances in grey matter density might help to characterise OCD by locating strong variations or dystrophies in the brain's structure. We will refer to this neuromarker as the VAR neuromarker from now on. Each VAR neuromarker for subject $\mathbf{x}_m$ and region S_g is computed as:

$$\mathrm{VAR}_m^{(g)} = \frac{1}{|S_g|} \sum_{i \in S_g} (x_{m,i} - \mathrm{AV}_m^{(g)})^2 \tag{4.3}$$

where $\mathrm{AV}_m^{(g)}$ is the AV neuromarker for $\mathbf{x}_m$ and brain region S_g.

4.2.4 SVM weighted grey matter probability

Finally, we can use the information provided by the linear SVM classifier to extract the relevant information of each brain region. We use the weights computed during the SVM training phase as a representation of each voxel in the same way that the classifier uses them to separate controls from patients.

As was seen in Chapter 2, a linear SVM classifier applied over the overall set of selected voxels S computes the output for a sample $\mathbf{x}$ as:

$$f(x) = \sum_{i \in S} w_i x_i + b \tag{4.4}$$

If we split the index set S into the different brain regions ($S = S_1 \cup S_2 \ldots \cup S_G$), (4.4) can be rewritten as:

$$f(x) = \sum_{g=1}^{G} \sum_{i \in S_g} w_i x_i + b \tag{4.5}$$

and each term of the inner summation would be summarizing the information of each region. We can now define the SVM weighted grey matter probability neuromarker for subject $\mathbf{x}_m$ and region S_g, which we will call the WE neuromarker, as:

$$\mathrm{WE}_m^{(g)} = \sum_{i \in S_g} w_i x_{m,i} \tag{4.6}$$

4.3 Neuromarker selection

The definition of these four types of neuromarkers gave us a much more intuitive set of features to work with. Instead of 43.000 voxels we now have, on average, 718 neuromarkers of four different types per subject. Furthermore, each neuromarker successfully represents a relatively large region of the brain, some of which are easily identifiable by psychiatrists as relevant, pathology related areas.

However, as was explained in Chapter 2, the process of feature selection can yield important benefits in terms not only of performance and generalization, but also in terms of ease of interpretation. For these reasons we decided to apply several feature selection algorithms to the neuromarkers defined above with the intention of arriving at a set of markers that is reduced and manageable enough to be useful to the psychiatric community.

After building the neuromarkers, a series of modifications to our notation

is required. The data matrix $\mathbf{X}$ now contains all the subjects, each one described by neuromarkers instead of voxels. D is thus no longer the total number of voxels, but the total number of neuromarkers. Also, the vector representing the values of the dth neuromarker (out of the total D) across all the M subjects will be represented as $\mathbf{n}^{(i)}$, where $\mathbf{n}^{(i)} = \left[n_1^{(i)}, n_2^{(i)} \dots n_M^{(i)}\right]$.

4.3.1 Variance based ranking

A quick glance over the neuromarker values reveals that some of them are constant over all subjects, regardless of whether they are patients or controls. This indicates that they will probably not be useful when discriminating patients from controls, therefore being irrelevant to OCD characterisation Therefore, a simple criterion to remove this redundancy is to apply a filtering process, ranking the neuromarkers according to their variance.

In this sense, the assumption is made that those neuromarkers that present a greater variance must be more relevant to OCD. The variance for each neuromarker, $\mathbf{n}^{(d)}$, was estimated using the unbiased variance estimator:

$$\hat{\sigma}^2(\mathbf{n}^{(d)}) = \frac{1}{M-1}\sum_{m=1}^{M}(n_m^{(d)} - \bar{\mathbf{n}}^{(d)})^2 \tag{4.7}$$

where $\bar{\mathbf{n}}^{(d)}$ is the average of the values of $\mathbf{n}^{(d)}$ across the M subjects.

4.3.2 Correlation based ranking

This criterion supposes that good neuromarkers should be highly correlated with the classification task. Thus, another straightforward selection procedure is to rank the neuromarkers according to their correlation with the classification labels [22].

The Pearson correlation coefficient between neuromarker $\mathbf{n}^{(d)}$ and label vector $\mathbf{y}$ is defined as follows:

$$\mathcal{R}(\mathbf{n}^{(d)}) = \frac{cov(\mathbf{n}^{(d)}, \mathbf{y})}{\sqrt{var(\mathbf{n}^{(d)})var(\mathbf{y})}} \tag{4.8}$$

where *cov* designates de covariance and *var* designates the variance. The sample correlation coefficient is an estimator for the Pearson correlation coefficient and it is given by:

$$\hat{R}(\mathbf{n}^{(d)}) = \frac{\sum_{m=1}^{M}(n_m^{(d)} - \bar{\mathbf{n}}^{(d)})(y_m - \bar{\mathbf{y}})}{\sqrt{\sum_{m=1}^{M}(n_m^{(d)} - \bar{\mathbf{n}}^{(d)})^2 \sum_{m=1}^{M}(y_m - \bar{\mathbf{y}})^2}} \tag{4.9}$$

where the bar notation stands for an average over the index m.

It is important to note that the Pearson correlation coefficient is only able to detect linear dependencies between a neuromaker and the labels.

Since greater values of R for a given neuromarker imply a stronger correlation between a neuromarker and the label vector, we interpret that those neuromarkers which produce greater correlation coefficients must be of greater relevance to OCD.

4.3.3 T-test based ranking

The third criterion applies a standard t-test [22] to analyse the statistical differences of the neuromarkers belonging to patient and control populations.

Specifically, we first separate the data set into patients and controls according to the labels. We then perform a two sample t-test of the hypothesis that the neuromarkers from the patient and control groups come from distributions with equal means. This is called the *null hypothesis*.

The t-test produces two results for each neuromarker: the test result, H, and the result's *p-value*. $H = 0$ indicates that the null hypothesis (that the means are equal) cannot be rejected at the 5% significance level. The *p-value* represents the probability of observing the given result, or one more extreme, by chance assuming that the null hypothesis is true. A small *p-value* casts doubt on the validity of the null hypothesis.

Since we are creating a ranked list of neuromarkes, the *p-value* is the most useful parameter to us. In this sense we interpret a small *p-value* for a neuromarker as an indication that it's distribution varies greatly from patients to controls and that it is therefore more relevant to OCD characterisation.

4.3.4 Forward-search by the Hilbert-Schmidt independence criterion

The correlation criterion analyses the linear relationships between features and labels. The strategy defined in this method extends this idea by measuring non linear relationships by means of the Hilbert-Schmidt independence criterion [20], [46].

Covariance allows us to measure linear relations between two variables:

$$\mathcal{C}_{AB} = \mathbb{E}_{AB}(\mathbf{A}\mathbf{B}^T) - \mathbb{E}_A(\mathbf{A})\mathbb{E}_B(\mathbf{B}^T) \tag{4.10}$$

This definition of covariance can be extended to a Hilbert space using kernel functions [18]:

$$\mathcal{C}_{AB} = \mathbb{E}_{AB}[(\phi(\mathbf{A}) - \mu_A) \otimes (\psi(\mathbf{B}) - \mu_B)] \tag{4.11}$$

where $\phi(\mathbf{A})$ and $\psi(\mathbf{B})$ are the kernel functions applied to $\mathbf{A}$ and $\mathbf{B}$ respectively, and $\mu_A = \mathbb{E}_A(\phi(\mathbf{A}))$ and $\mu_B = \mathbb{E}_B(\psi(\mathbf{B}))$.

The Hilbert-Schmidt independence criterion (HSIC) is provided by the norm-2 of the covariance obtained in the Hilbert space, $||\mathcal{C}_{AB}||^2_{HS}$, which can be expressed in kernel terms as:

$$\text{HSIC}(\mathbf{A}, \mathbf{B}) = \frac{1}{m^2}\text{Tr}(\tilde{K}_A\tilde{K}_B) \tag{4.12}$$

where $\tilde{K}_A$ and $\tilde{K}_B$ are the centred kernel matrices associated with the variables $\mathbf{A}$ and $\mathbf{B}$.

If a linear kernel function is used, calculating HSIC between two variables is equivalent to calculating their correlation. However, if a non-linear kernel function is used, non linear relations between the variables will be computed. In the case of our work we used a Gaussian kernel function. When the Gaussian kernel function is applied to a set of neuromarkers, it yields a kernel matrix, $\mathbf{K}$, for which each individual element, $k_{i,j}$, is computed as

follows:

$$k_{i,j} = \exp\left(-\frac{||\mathbf{n}_i - \mathbf{n}_j||^2}{2\sigma^2}\right) \tag{4.13}$$

where $\mathbf{n}_i$ and $\mathbf{n}_j$ are two neuromarkers from the set.

As with the correlation criterion, we interpret that larger values for the HSIC test imply a higher relation between neuromarkers and labels. In this case both a standard ranking and a forward-search algorithm were implemented, but only the forward search approach provided significant results. As was explained in Section 2.2, forward-search algorithms don't produce a ranking of features but rather a series of *nested subsets* in order of relevance. Figure 4.1 depicts the forward-search HSIC selection algorithm.

Since this selection algorithm is computationally very intensive, the first iterations were programmed so that they eliminated the 10 least relevant features each time until only 100 features were left. From this point on we continued by eliminating only one feature per iteration.

4.3.5 Recursive feature elimination

Recursive feature elimintaion (RFE) was first proposed in [21] as an instance of backward feature elimination applied to SVM classifiers. Since RFE selects features according to the classification margin provided by the SVM classifier, it does not separate the learning process from the feature selection process. For this reason, it falls into the category of embedded feature selection algorithms [23].

RFE aims at finding the subset of features that is able to provide the largest classification margin in an SVM classifier. To achieve this goal it starts by training an SVM with all the features and then analysing which of the features can be eliminated while producing the smallest variation in the classification margin. This process is repeated, each time eliminating one feature, until no features are left. Thus, in each iteration the algorithm trains an SVM classifier using a smaller subset of features than in the previous iteration. Figure 4.2 describes the RFE algorithm in detail.

If, as is our case, the classification process is using a linear SVM, the absolute value of the weight vector for a given feature, $|w_d|$, can be used as

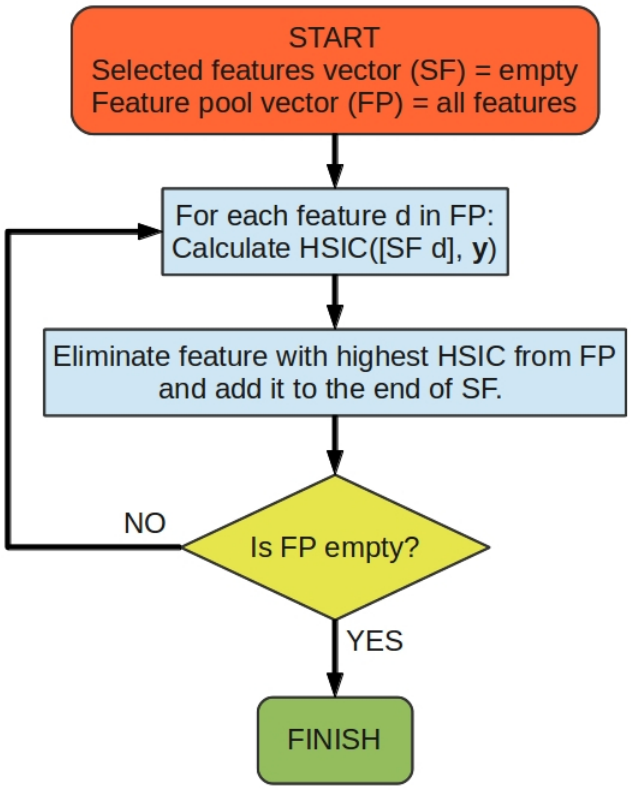

Figure 4.1: Flow-chart depicting the forward-search HSIC algorithm. The notation HSIC([SF f], $\mathbf{y}$) means that we calculate HSIC between the features listed in SF with feature f appended and the label vector $\mathbf{y}$.

the relevance criterion, where $\mathbf{w}$ is the weight vector obtained from the SVM classifier and is calculated as (see Section 2.3):

$$\mathbf{w} = \sum_{m=1}^{M} \alpha_m y_m \mathbf{x}_m \tag{4.14}$$

In each iteration, the feature that corresponds with the lowest w_d will be eliminated since it is the one that produces the smallest variation in the classification margin.

As with the previous feature selection criteria, this recursive elimination process will provide *nested subsets* of neuromarkers, each being more discriminating than the previous one.

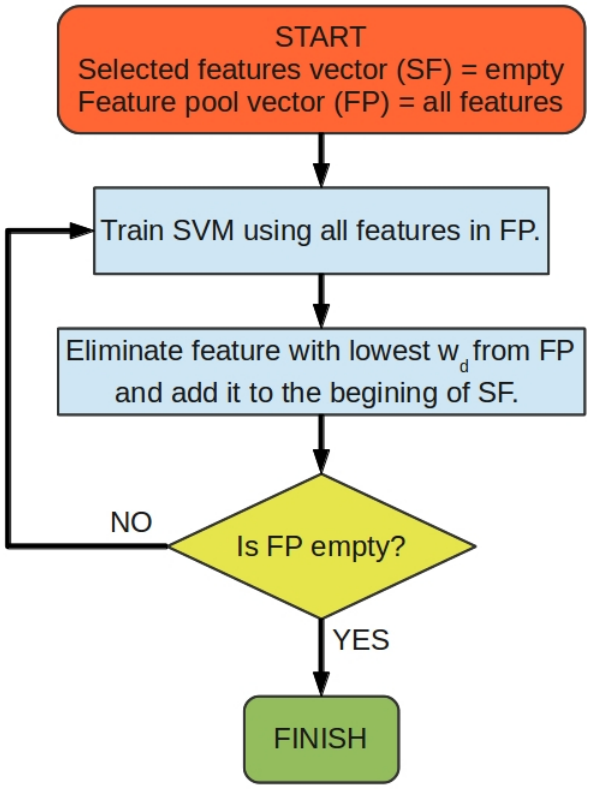

Figure 4.2: Flow-chart depicting the RFE algorithm.

As with the previous algorithm, the first iterations of the RFE strategy were programmed so that they eliminated 10 features at a time until only 100 were left. From then on it proceeded by eliminating one at a time.

Chapter 5

Experiments

This chapter shows and explains the results obtained after applying the methods described in Chapter 4.

It starts by describing the double leave-one-out algorithm, which is the specific validation and testing strategy that was used.

It then analyses the performance of each neuromarker type paired with each feature selection strategy and determines which is the most effective combination. It also shows the evolution of the classification test error with the number of selected neuromarkers.

Finally, it maps the most relevant neuromarkers of the winning subset to a brain template and renders them so that they may be visually analysed. To this end, input from the medical personnel of the Bellvitge University Hospital is included.

The algorithms were designed and tested on a ThinkPad L512 laptop running Matlab 8.01 (The MathWorks Inc, Natick, Mass) on a Linux Mint (www.linuxmint.com) operating system. The Matlab SVM classification library that was used is libSVM (www.csie.ntu.edu.tw/c̃jlin/libsvm). Data processing, which was very computationally intensive, was performed on a Fura computer cluster where each node was running Matlab 7.8 on a Linux Gentoo (www.gentoo.com) operating system. The results were brought back to the ThinkPad laptop for analysis and visualisation purposes. MRI scans were processed using MRIcro (www.mricro.com) for the 2-D renders and

MRIcroGL (www.cabiatl.com) for the 3-D renders. Both these programmes are open-source.

5.1 Validation and testing strategy: the double leave-one-out algorithm

For our experiments, a great deal of validation had to be performed to determine the optimal number of features for each neuromarker type and selection algorithm. After this validation process is done, the method's generalisation capabilities need to be tested. In Section 2.2 we emphasised the importance of employing a proper training and validation strategy. Because the amount of subjects we have in this case is extremely small, it would be inadequate to simply split the training data into training and validation subsets because we would increase the severity of the small sample problem even more.

Other validation strategies like K-fold cross-validation employ a different tactic: instead of simply splitting the training data into two subsets, it randomly splits it into K different subsets; it then uses a single subset for validation and the rest of the subsets for training; the process is repeated K times, with each iteration using a different subset for validation; finally, the K validation results can be averaged (or otherwise combined) to produce a single estimation. There is no risk of overfitting since in each iteration of the process the training and validation subsets are composed of different samples. Figure 5.1 depicts the K-fold validation algorithm.

When $K = M$, M being the total number of subjects in the original training set, the K-fold cross-validation goes through exactly M iterations of the validation process. In each iteration only one of the subjects is used for validation while the remaining $M - 1$ subjects are used as the training data. This is known as the *leave-one-out* (LOO) cross-validation algorithm.

The LOO cross-validation strategy serves our purposes well since it maximises the amount of data available to us for training. However, we still need a testing data set to calculate our classification error rates. To overcome this problem, a *double LOO* (2LOO) strategy was employed. This idea extends

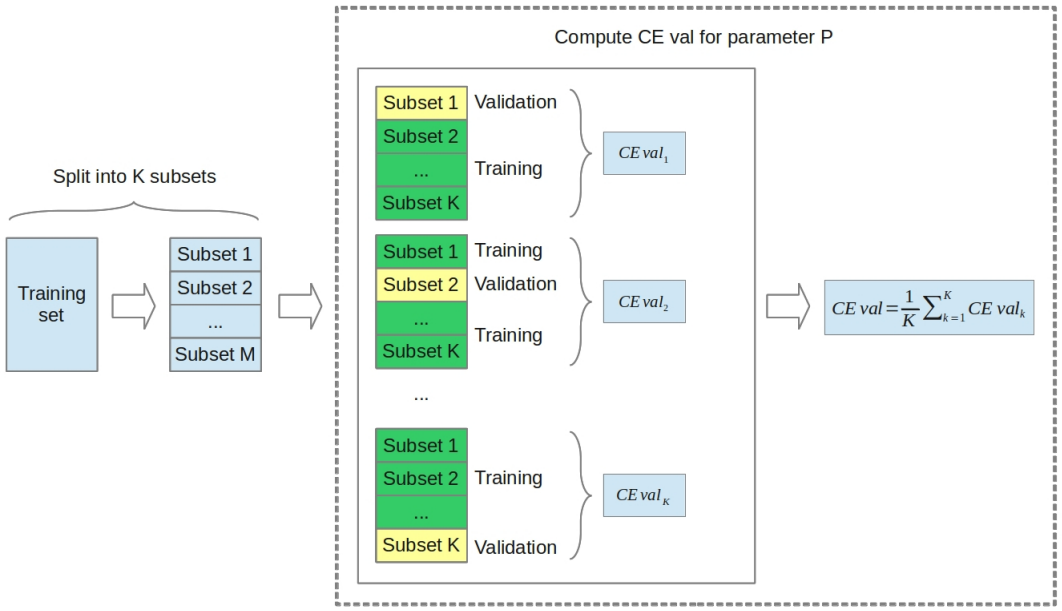

Figure 5.1: The K-fold cross-validation strategy.

LOO cross-validation by nesting one LOO algorithm inside another. An intuitive description of this concept is to imagine that we have M independent universes, one for each subject in the original data. In each of these universes we have a single test subject and $M-1$ training subjects. We then validate a parameter by executing the nested LOO cross-validation algorithm with the $M-1$ training subjects, where each iteration of the nested LOO will train with $M-2$ subjects; we then calculate the predicted label for the single test subject using the validated parameter. The classification error will be either 0 or 1 depending on whether the label was correctly or incorrectly predicted. If we repeat this process for each of our M independent universes, we will end up with M different classification errors. The average of these independent classification errors will be the global classification test error for our problem. Figure 5.2 illustrates this process.

In order to create our feature rankings and nested subsets, we applied the 2LOO algorithm to each of the 172 subjects in our data. For each subject we ranked the features through LOO cross-validation using the remaining 171 subjects and following the selection criteria described in Chapter 4. We

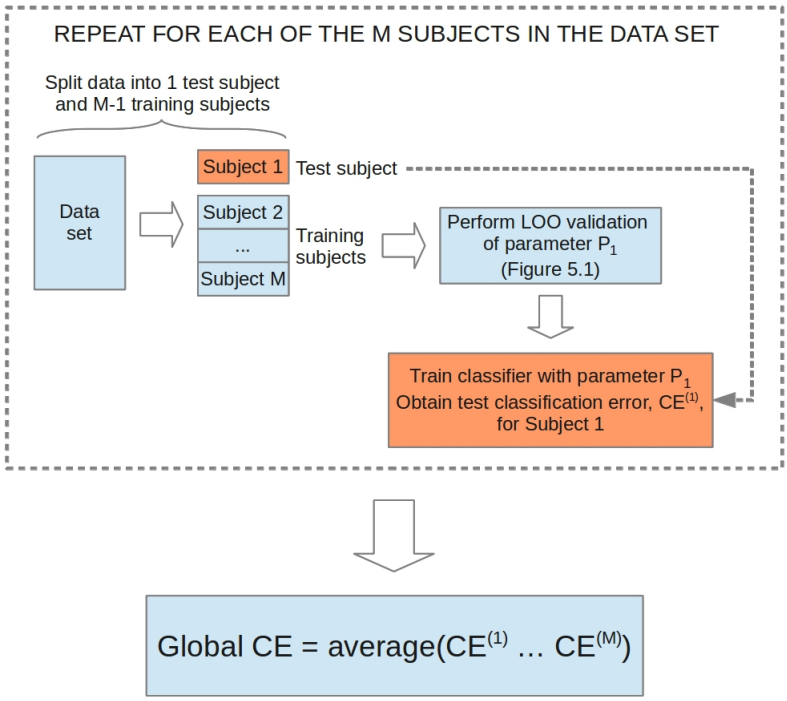

Figure 5.2: The 2LOO validation and test strategy.

then predicted each test subject's test label using the subset of features that provided the lowest validation error in its 2LOO universe. To obtain the global classification error we averaged the prediction test errors obtained in each of the 2LOO universes.

5.2 Performance analysis

In this section we analyse the usefulness of each of the neuromarkers from Section 4.2 and the extent to which we can reduce their number by means of the feature selection strategies described in Section 4.3. To this end we shall consider a neuromarker to be useful if we can maintain a classification error that is similar to the 26.2% obtained with the T-BS process from Chapter 3.

Table 5.1: Analysis of the 2LOO classification errors (CE) and average number of neuromarkers (# NM) obtained by selection criteria and neuromarker type.

		AV	**VAR**	**ACC**	**WE**
All neuromarkers	CE (%)	35.47	49.42	33.14	28.49
	# NM	718	718	718	718
Variance ranking	CE (%)	34.30	50.58	36.63	28.49
	# NM	96.22	144.81	40.02	38.84
t-Test selection	CE (%)	34.88	52.91	38.37	32.56
	# NM	258.25	198.75	286.65	525.83
RFE	CE (%)	36.05	51.16	32.56	30.23
	# NM	132.88	245.03	224.45	60.26
HSIC-Test ranking	CE (%)	38.37	47.09	40.12	31.98
	# NM	96.81	148.20	54.76	48.41
Correlation ranking	CE (%)	40.12	50.00	40.70	32.56
	# NM	575.40	435.61	513.93	527.51

Table 5.1 illustrates the effectiveness of our neuromarker types, paired with each selection strategy, at characterising OCD by means of the classification error and the number of neuromarkers that yield said error. For comparison, the first row shows the error rate obtained without applying any selection algorithm.

Overall, the most capable neuromaker is by far the WE type. Most selection criteria converge at errors of around 30% when applied with it. Moreover, the number of relevant features needed to characterise the pathology using this neuromarker is of around 50 with the variance ranking, HSIC ranking and RFE methods.

Both the ACC and AV neuromarkers manage to characterise OCD fairly well, obtaining error rates smaller than 40% with reduced numbers of neuromarkers. The VAR neuromarker type, on the other hand, performs very poorly since it doesn't seem to accurately discriminate the disorder at all (a classification error of 50% in a binary problem means that the model is performing no better than a random classifier).

Specifically, the most effective criterion overall is the WE neuromarker

paired with the variance ranking selection strategy. This combination produces a global test error of 28.49%, which is only slightly greater than the error obtained before the neuromarker construction and selection process, while the number of features it employs is one order of magnitude smaller: only 38.8 on average over the 172 subjects as opposed to the 718 original brain regions obtained from the T-BS process. It is important to remember that the initial problem had a dimensionality of 500.000 variables in the form of voxels versus a very scarce number of subjects. We have now managed to greatly reduce the problem by employing only around 40 neuromarkers on average, all the while keeping the classification error below 30%.

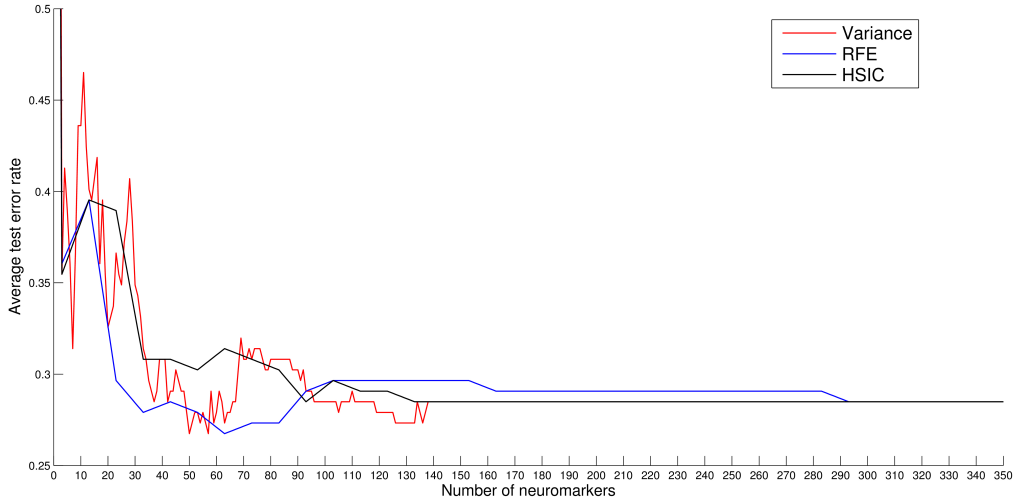

Figure 5.3: Evolution of the average test error rate with subset size for the WE neuromarker.

Figure 5.3 depicts the evolution of the classification test error rate with the number of neuromarkers used for testing for the WE neuromarker and the three most successful selection strategies. It can be seen that the error rates show very little variation as we decrease the number of neuromarkers. The error rates start to vary significantly once we start training with less than 150 neuromarkers. This points to high redundancy or low relevance in the data up until that point. Only after we have decreased the subset size to under 35 neuromarkers does the error rate begin to steadily increase. As

is expected, when the number of neuromarkers used is high, the error rate converges to the values obtained when no selection is applied.

5.3 Visualizing neuromarkers

Given the results of the previous section, we will now analyse the relevance and neuroanatomical position of the most important WE neuromarkers selected by the variance ranking method.

Due to the fact that we are employing the 2LOO validation and test strategy over the 172 study subjects, we have obtained 172 different subsets of neuromarkers with an average size of 38.84. In order to obtain a single subset that is easy to interpret, we have merged the 172 subsets into a single one of size 59, where each of its neuromarkers appears in at least one of the original 172 subsets. Note that these neuromarkers present varying consistencies in the voxels they contain over the 172 iterations, meaning that some of these voxels appear in every iteration while others in just a few.

To establish the relevance of each neuromarker we have applied a simple algorithm that resembles a single iteration of the wrapper feature selection method described in Section 2.2. We obtain the classification error rates when each of the 59 neuromarkers is eliminated from the training set. These error rates are then compared to the error rate obtained when all 59 neuromarkers are used in the training process. The neuromarkers whose elimination from the training set produce a greater variation in the classification error will be ranked higher.

Table 5.3 shows these classification error rate deviations for the most important neuromarkers, that is, those which cause a significant classification increment when they are not used during training. The analysis is completed by associating each of these neuromarkers with the most significant MNI neuroanatomical regions in which they exist [50] (those regions whose voxel consistency is greater than 50% over the 172 subjects).

The neuromarkers can be remapped to their component voxels and exported as NIfTI files, which is the standard sMRI visualisation format. Figures 5.4 - 5.16 show the localization of the thirteen most important neuro-

markers within the brain, with color intensity indicating the consistency of each voxel across the study subjects (bright yellow means that a particular voxel was selected in all the iterations, while redder tones imply lower consistency). In these figures, the top left image is a vertical cut seen from the rear of the brain, the top right image is a vertical cut seen from the right hand side of the brain, bottom left is a horizontal cut viewed from the top of the brain and, finally, bottom right is a 3-D visualisation of the entire brain in which the appropriate neuromarker has been highlighted.

At this stage, the data was sent back to Dr. Carles Soriano Mas of the Department of Psychiatry of the Bellvitge University Hospital in Barcelona. After his analysis it was determined that the five most relevant neuromarkers to OCD (Figures 5.4 - 5.8) are located in the frontal, temporal and parietal lobes. Three of them appear in regions traditionally associated with the disorder, such as the orbitofrontal cortex (right inferior frontal and middle frontal gyri) and the striatum (putamen and globus pallidum, extending to the adjacent insular cortex). Such regions are part of the distributed cortico-striatal circuits know to be involved in OCD pathophysiology [24]. Specifically, while striatal regions seem to be hyperactive (their volume appears increased in patients), prefrontal areas seem to be hypoactive (their volume appears decreased in patients) and inefficient in regulating enhanced striatal activity, which leads to the development of the repetitive and ritualized behaviors characteristic of OCD.

Other regions present in our neuromarkers, such as the superior temporal and supramarginal gyri, have been less frequently associated with the disorder, although they are also connected to subcortical striatal regions and thus may also be considered as part of the extended cortico-striatal circuitry. Indeed, the role of the parietal cortex (i.e., supramariginal gyrus) in striatal regulation and the importance of such parieto-striatal connectivity for OCD has already been incorporated in more recent neurobiological models of the disease [36].

Table 5.2: Neuroanatomical analysis for the most important neuromarkers

NM ranking	Δ CE (%)	MNI ROIs
1	6.39	Temporal Sup-Mid L
2	5.81	Frontal Inf Tri-Orb R; Insula R
3	4.65	Insula L; Putamen L; Pallidum L
4	4.65	Parietal Inf L; SupraMarginal L
5	3.49	Frontal Sup-Mid R
6	2.9	Calcarine L-R; Lingual L; Precuneus L-R; Cerebelum 6 L; Vermis 4-5-6
7	2.9	Olfactory L-R; Frontal Med Orb L; Rectus L; Cingulum Ant R; Lingual L-R; Occipital Inf R; Fusiform L-R; Precuneus L; Caudate L; Pallidum L; Thalamus L-R; Temporal Inf R; Cerebelum Crus-1-3-4-5-6-7b-9-10 L-R; Vermis 1-2-3-4-5-7-10
8	2.9	Temporal Sup-Mid R
9	2.9	Frontal Inf Oper-Tri R; Insula R; Putamen R; Pallidum R; Heschl R; Temporal Sup-Pole Sup R
10	2.9	Precentral R; Frontal Mid R; Postcentral R; Parietal Inf R; SupraMarginal R
11	2.9	Parietal Inf R; SupraMarginal R; Angular R; Temporal Sup R
12	2.9	Frontal Mid L
13	2.9	Cerebelum Crus2-7b-8 R
14	2.32	Lingual R
15	2.32	Fusiform L; Temporal Inf L
16	2.32	Frontal Sup-Mid L
17	2.32	Frontal Med Orb L-R; Rectus L-R
18	2.32	Frontal Sup Orb L; Rectus L
19	1.74	Cuneus L; Parietal Sup L; Precuneus L
20	1.16	Occipital Inf L; Parietal Sup-Inf L; SupraMarginal L; Angular L; Temporal Sup-Mid L
21	1.16	Temporal Sup-Mid-Inf R
22	1.16	Temporal Mid L
23	0.58	Precentral; Frontal Sup-Mid L
24	0.58	Frontal Sup Medial L-R; Cingulum Ant L-R
25	0.58	Precentral L; Frontal Mid L
26	0.58	Hippocampus L-R
27	0.58	Occipital Sup-Mid R

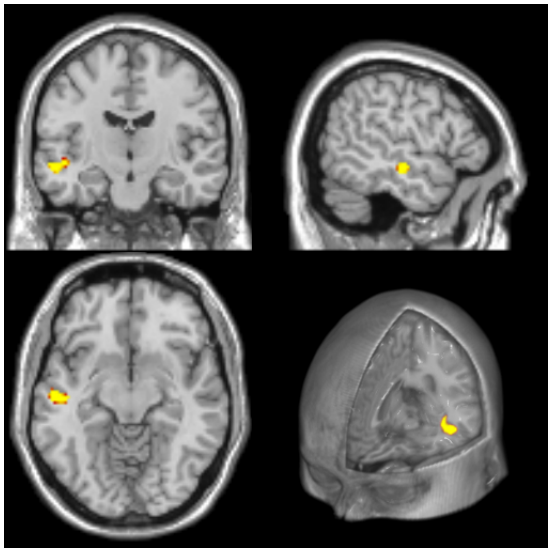

Figure 5.4: Position of the first ranked neuromarker: Temporal Sup-Mid Left.

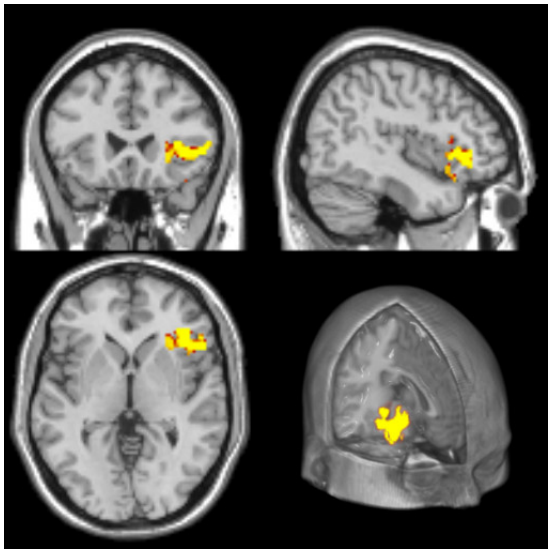

Figure 5.5: Position of the second ranked neuromarker: Frontal Inf Tri-Orb Right; Insula Right.

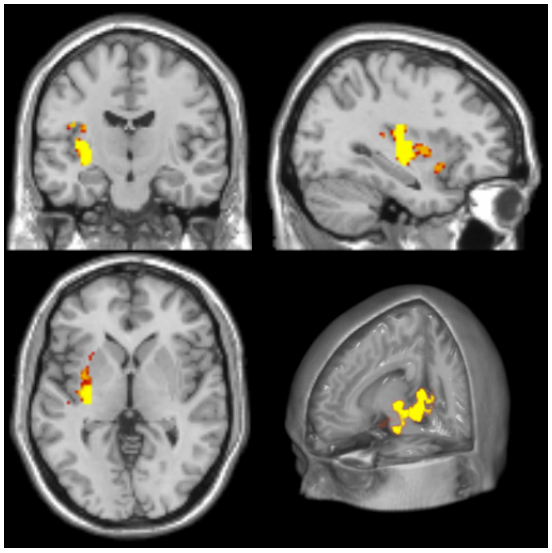

Figure 5.6: Position of the third ranked neuromarker: Insula Left; Putamen Left; Pallidum Left

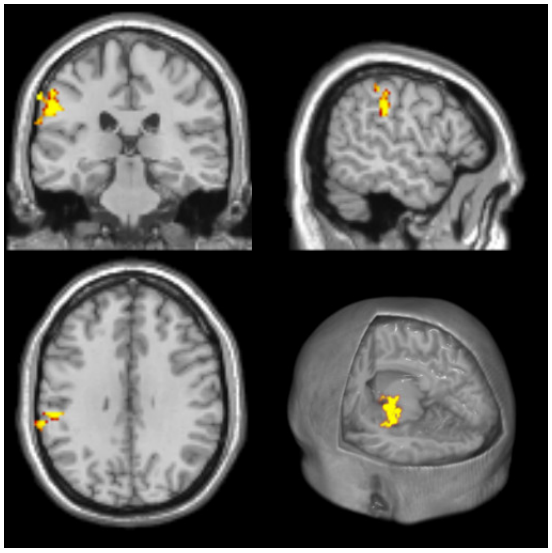

Figure 5.7: Position of the fourth ranked neuromarker: Parietal Inf Left; SupraMarginal Left.

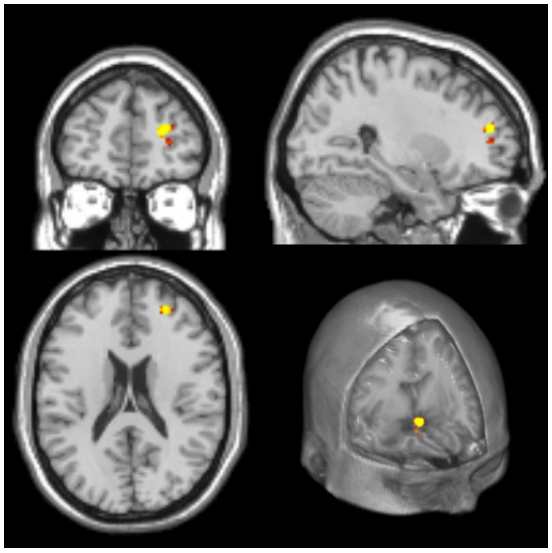

Figure 5.8: Position of the fifth ranked neuromarker: Frontal Sup-Mid Right.

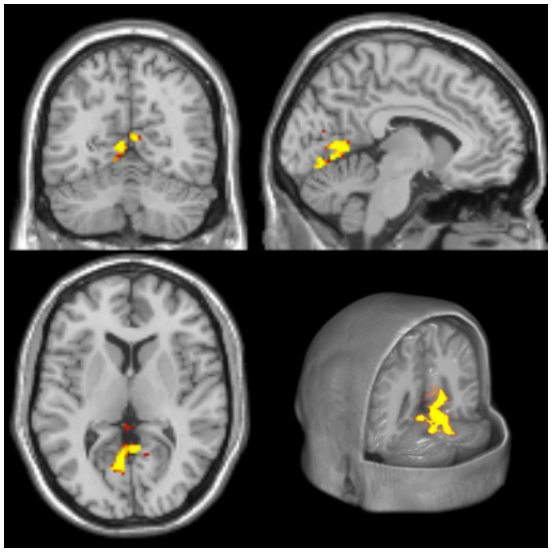

Figure 5.9: Position of the sixth ranked neuromarker: Calcarine Left-Right; Lingual Left; Precuneus Left-Right; Cerebelum 6 Left; Vermis 4-5-6

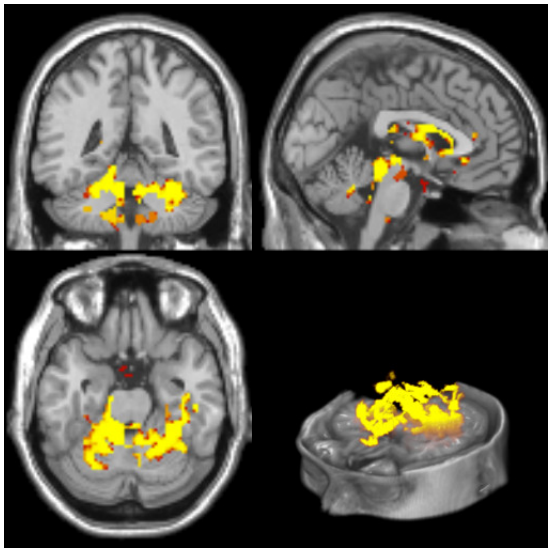

Figure 5.10: Position of the seventh ranked neuromarker: Olfactory L-R ; Frontal Med Orb L; Rectus L; Cingulum Ant R; Lingual L-R; Occipital Inf R; Fusiform L-R; Precuneus LCaudate L; Pallidum L; Thalamus L-R; Temporal Inf R; Cerebelum Crus L-R; Vermis.

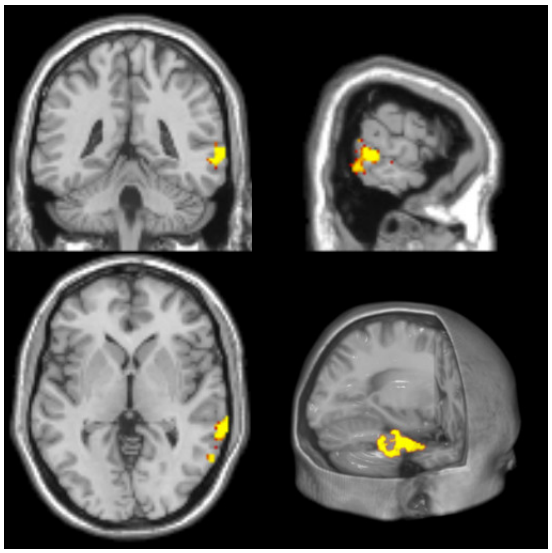

Figure 5.11: Position of the eighth ranked neuromarker: Temporal Sup-Mid Right.

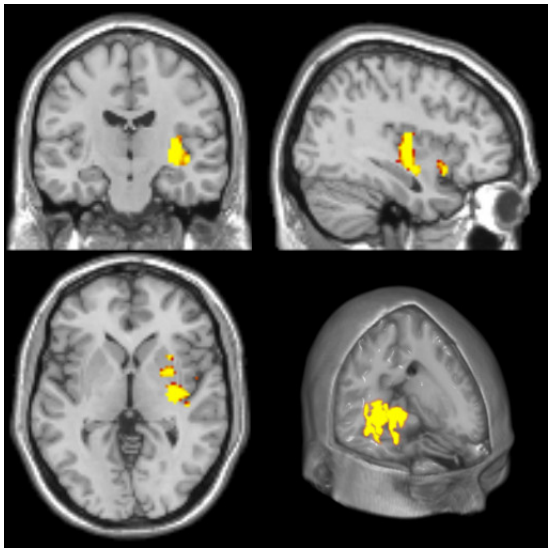

Figure 5.12: Position of the ninth ranked neuromarker: Frontal Inf Oper-Tri R; Insula R; Putamen R; Pallidum R; Heschl R; Temporal Sup-Pole Sup R.

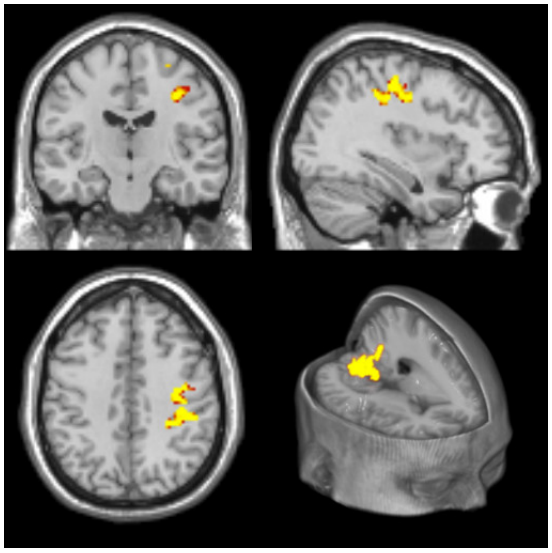

Figure 5.13: Position of the tenth ranked neuromarker: Precentral R; Frontal Mid R; Postcentral R; Parietal Inf R; SupraMarginal R.

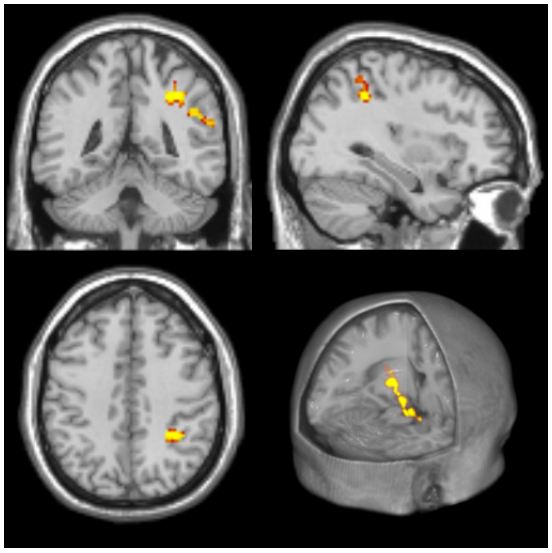

Figure 5.14: Position of the eleventh ranked neuromarker: Parietal Inf R; SupraMarginal R; Angular R; Temporal Sup R.

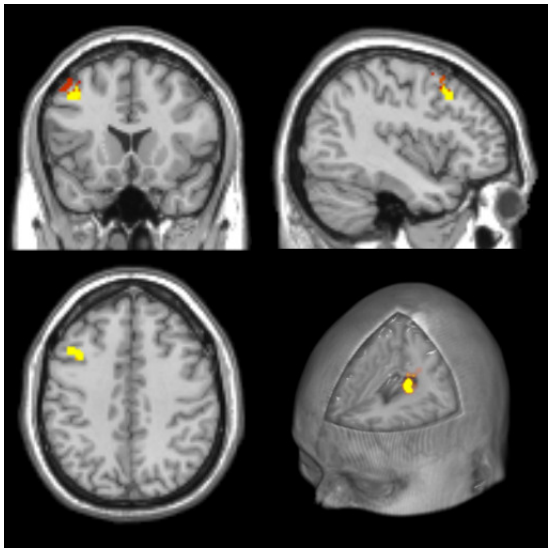

Figure 5.15: Position of the twelfth ranked neuromarker: Frontal Mid L.

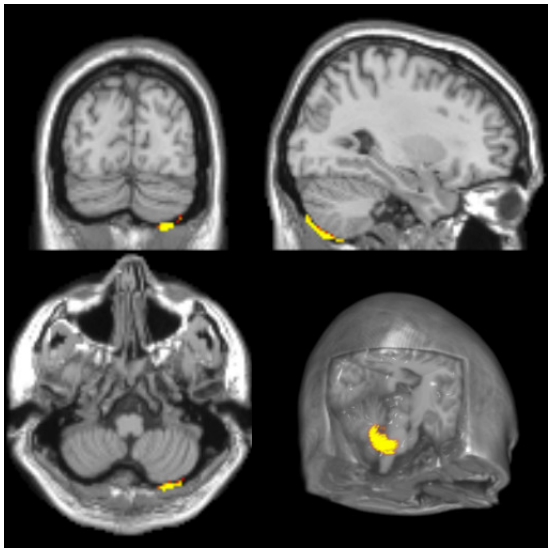

Figure 5.16: Position of the thirteenth ranked neuromarker: Cerebelum Crus2-7b-8 R.

Chapter 6

Conclusions and future lines of investigation

This thesis set out to establish a framework to automatically obtain a set of neuromarkers capable of characterizing OCD. To this end, it follows on from the conclusions reached by the studies presented in [40], where an average of 43.000 voxels had been identified as being relevant to the classification task of OCD afflicted brains. Although this is a much better analysis scenario for the psychiatric community than the initial 500.000 voxels produced by a standard sMRI, it is still a very abstract and difficult characterisation of the brain of an average OCD patient.

First we needed to transform this cumbersome and unfriendly amount of data into a series of intuitive measurements associated with the 718 separate brain regions, on average per subject, that contained the 43.000 voxels. These measurements, defined as neuromarkers, would then be tested for their relevance by analysing their performance when used as input data for the task of discerning healthy brains from unhealthy ones.

The presented work analyses four different kinds of neuromarkers candidates:

- The AV neuromarker candidate is a measurement of the average grey matter density of each region.
- The ACC neuromarker measures the accumulated grey matter density

of each region.

- The VAR neuromarker represents the variance of the grey matter density contained in each region.
- The WE neuromarker is a measurement of the SVM linear classifier weights associated to the voxels contained in each region.

Once the data is processed we are left with an average of 718 possible neuromarkers per subject, each representing a differentiated brain region with a single measurement instead of hundreds of voxels. An initial test of the viability of these measurements as neuromarker candidates yields very positive results: when used to train and test a classifier, three of them prove to be effective at characterising OCD, with test error rates of under 40% for the AV, ACC and WE neuromarkers.

Specifically, the WE neuromarker obtains a classification test error rate of 28.49%, which is only very slightly greater than the 26.2% error rate obtained when classifying with the 43.000 voxels. More importantly, these neuromarkers are obtaining much better error rates than the approximately 40% that was obtained when testing a classifier with all the 500.000 voxels produced by an sMRI brain scan. This means that a lot of redundant or irrelevant information has been eliminated.

The next step is to test whether or not all of the 718 neuromarkers are useful or, on the other hand, a further reduction aimed at finding more relevant neuromarkers can be achieved. It must be remembered that our goal is to define a limited number of neuromarkers that are manageable and useful to medical personnel. In this regard, the fewer relevant neuromarkers we end up with, the better our system will be.

To achieve this, we apply a series of state of the art feature selection methods to our neuromarkers. The results of these selection processes are very positive. Experimental results reveal that the WE neuromarker candidate in combination with a selection based on its variance is able to provide a subset of no more than 50 values that are easy to interpret and handle by the psychiatric community. Furthermore, this is achieved while retaining a

test classification error which is only slightly greater than the error obtained when classifying with the 43.000 voxels discovered in [40]. We have now gone from 500.000 voxels initially produced by an sMRI brain scan to just under 50 neuromarker candidates that represent as many relevant brain regions with a single value.

In order for this neuromarker candidate to be confirmed as such it needs to meet the definition for a neuromarker presented in Chapter 1. We consider that the the WE neuromarker candidate adequately fulfils the task of quantifying a neuroanatomical characteristic associated with a pathology.

Future lines of work will be focused on studying further compliance with the neuromarker definition:

- Can these neuromarkers also be used to analyse the patient's evolution?
- Can these neuromarkes help in detecting a pathology's subtype?
- Can these neuromarkers provide aid in the prescription process?

In order to pursue these goals, more data will be needed. Multi-class classification analysis can be done if we can obtain clearly differentiated sMRI brain scans of separate endophenotypes of OCD, or of patients in different stages of illness evolution.

Furthermore, we also intend to extend this framework to other pathologies that could benefit from being characterized by neuromarkers. Structural and functional brain anomalies have already been pointed out to be at least partly responsible for Alzheimer's disease [53] and schizophrenia [12].

Extending our research to these ailments could prove to be extremely useful in testing the effectiveness of our methods. Eventually, our work could provide society with a very powerful and useful tool in the diagnosis and treatment of mental illness.

Chapter 7

Research project budgets and planning

In the first section of this chapter we will discuss project planning and layout. The second section deals with cost justification and budgets.

7.1 Project planning

The initial stage of the project consisted on an intensive tutor-student information transfer. Many concepts that go beyond the scope of the student's bachelor degree, especially those concerning ML, needed to be thoroughly understood in order to give the student the necessary tools to be able to develop and test the required methods. To further set the foundations of the student's understanding of ML, the 10 week ML online course by Proff. Andrew Ng from Standford University, through the Coursera online learning platform, was taken.

The next stage was a familiarisation with the work previous to this thesis. Apart from the work presented in [40], numerous articles and text books on themes ranging from ML to neuroimaging were studied. During this stage, the software tools that were going to be used, specifically the libSVM classifier library, were also studied.

The third stage was the design and testing of all the algorithms and

methods. Many refinements and fine-tuning processes were required in order to adapt the algorithms to the complex dimensionality of the database that was being used.

Finally, the methods were applied to the data and the results were analysed by the student, the tutor and the medical personnel at the Bellvitge Hospital. A 15 page paper on the subject was presented to the 2014 European Congress of Machine Learning (ECML/PKDD 2014; paper acceptance rate of 23.8% for 2014) with the student as the main author. The paper has been accepted for presentation.

During the entire project, the student worked in close collaboration with the tutor, holding weekly meetings to oversee progress and maintaining daily communication via e-mail.

Figure 7.1 shows a comprehensive list of the primary and secondary stages that comprise the project. Figure 7.2 shows a Gantt graph depicting the evolution of the project.

WBS	Name	Start	Finish	Duration
1	Project planning	Sep 16	Sep 17	2d
2	Conceptual transfer	Sep 18	Oct 15	20d
3	Tool familiarisation	Oct 16	Nov 12	20d
4	Study of previous work	Oct 16	Nov 12	20d
5	Data analysis	Nov 13	Nov 19	5d
6	NM building methods dev. and test	Nov 20	Dec 17	20d
7	NM selection methods dev. and test	Nov 20	Jan 7	35d
8	NM extraction	Dec 18	Jan 14	20d
9	NM selection	Jan 15	Mar 4	35d
10	Performance analysis	Mar 5	Mar 18	10d
11	NM visualisation	Mar 19	Mar 21	3d
12	Medical feedback	Mar 24	Mar 28	5d
13	ECML paper	Mar 31	Apr 18	15d
14	Thesis summary	Apr 21	Jun 13	40d

Figure 7.1: Project task list. The project started on the sixteenth of September, 2013.

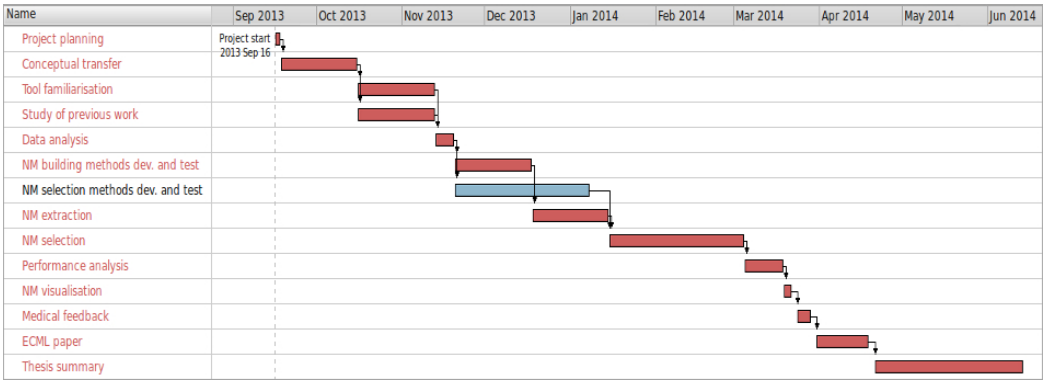

Figure 7.2: Project Gantt graph. Tasks coloured in red are time-critical tasks.

7.2 Project budgets

This section presents a justification of the overall project costs. We analyse both personnel costs as well as material resource costs to finally arrive at a global budget figure.

7.2.1 Personnel costs

Three people have participated in this project's development:

- The student, considered a junior engineer.
- The tutor, a senior researcher.
- Doctor Carles Soriano Mas, involved only in the results analysis phase.

From the development times described in Figure 7.1, Table 7.1 shows the number of hours that each person has dedicated to the project and the associated costs.

7.2.2 Material resources costs

During the development of the project several material resources have been used. These are listed with their associated costs in Table 7.2.

Table 7.1: Personnel costs.

Personnel	**Work hours**	**Hourly Rate** (€/h)	**Total** (€)
Junior engineer	980	12	11.760
Senior researcher	155	23	3.565
Medical personnel	20	30	600
		TOTAL	15.925

Table 7.2: Material resources costs. Amortisation is yearly.

Concept	**Quantity**	**Price** (€/unit)	**Amort.**	**Total** (€)
MRI brain scan	172	150	50%	12.900
Matlab License	1	2.000	100%	2.000
ThinkPad L512	1	850	25%	212,5
Computer cluster	100 nodes	40 per node	20%	800
			TOTAL	15.912,5

7.2.3 Total project budget

The overall budget for the project is presented in Table 7.3.

Table 7.3: Overall budget.

Concept	**Total** (€)
Personnel costs	15.925
Material resource costs	15.912,5
Total costs	**31.837,5**
VAT (21%)	6.685,9
Total Budget	**38.523,4**

Bibliography

[1] *Statistical Parametric Mapping. The Analysis of Functional Brain Images*. Elsevier, 2007.

[2] Edoardo Amaldi and Viggo Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1):237–260, 1998.

[3] John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

[4] John Ashburner and Karl J Friston. Voxel-based morphometry—the methods. *Neuroimage*, 11(6):805–821, 2000.

[5] Murad Atmaca, Hanefi Yildirim, Huseyin Ozdemir, Ertan Tezcan, and A Kursad Poyraz. Volumetric MRI study of key brain regions implicated in obsessive–compulsive disorder. *Progress in neuro-psychopharmacology and Biological Psychiatry*, 31(1):46–52, 2007.

[6] J Bi, K Bennett, M Embrechts, C Breneman, and M Song. Dimensionality reduction via sparse support vector machines. *JMLR*, 3:1229–1243, 2003.

[7] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[8] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the*

fifth annual workshop on Computational learning theory, pages 144–152. ACM, 1992.

[9] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[10] Hans C Breiter and Scott L Rauch. Functional MRI and the study of OCD: from symptom provocation to cognitive-behavioral probes of cortico-striatal systems and the amygdala. *Neuroimage*, 4(3):S127–S138, 1996.

[11] Cameron S Carter, Angus W MacDonald, Laura L Ross, and V Andrew Stenger. Anterior cingulate cortex activity and impaired self-monitoring of performance in patients with schizophrenia: an event-related fMRI study. *American Journal of Psychiatry*, 158(9):1423–1428, 2001.

[12] Eduardo Castro, Manel Martínez-Ramón, Godfrey Pearlson, Jing Sui, and Vince D Calhoun. Characterization of groups using composite kernels and multi-source fMRI analysis data: application to schizophrenia. *Neuroimage*, 58(2):526–536, 2011.

[13] ARNOLD M COOPER and Robert Michels. Diagnostic and statistical manual of mental disorders. *American Journal of Psychiatry*, 138(1):128–129, 1981.

[14] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[15] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[16] Jorge De La Calleja and Olac Fuentes. Machine learning and image analysis for morphological galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 349(1):87–93, 2004.

[17] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[18] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007.

[19] Catriona D Good, Ingrid S Johnsrude, John Ashburner, Richard NA Henson, KJ Fristen, and Richard SJ Frackowiak. A voxel-based morphometric study of ageing in 465 normal adult human brains. In *Biomedical Imaging, 2002. 5th IEEE EMBS International Summer School on*, pages 16–pp. IEEE, 2002.

[20] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005.

[21] I Guyon, J Weston, S Barnhill, and V Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[22] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[23] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and L Zadeh. Feature extraction. *Foundations and applications*, 2006.

[24] Ben J Harrison, Carles Soriano-Mas, Jesus Pujol, Hector Ortiz, Marina López-Solà, Rosa Hernández-Ribas, Joan Deus, Pino Alonso, Murat Yücel, Christos Pantelis, et al. Altered corticostriatal functional connectivity in obsessive-compulsive disorder. *Archives of general psychiatry*, 66(11):1189–1200, 2009.

[25] John-Dylan Haynes and Geraint Rees. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature neuroscience*, 8(5):686–691, 2005.

[26] David W Hosmer Jr and Stanley Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2004.

[27] Andriana Iankova. The Glasgow coma scale clinical application in emergency departments. *emergency nurse*, 14(8):30–35, 2006.

[28] Douglas B Kell and Ross D King. On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. *Trends in Biotechnology*, 18(3):93–98, 2000.

[29] Stefan Klöppel, Ahmed Abdulkadir, Clifford R Jack Jr, Nikolaos Koutsouleris, Janaina Mourão-Miranda, and Prashanthi Vemuri. Diagnostic neuroimaging across diseases. *Neuroimage*, 61(2):457–463, 2012.

[30] William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. APACHE II: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

[31] Stephen LaConte, Stephen Strother, Vladimir Cherkassky, Jon Anderson, and Xiaoping Hu. Support vector machines for temporal classification of block design fMRI data. *NeuroImage*, 26(2):317–329, 2005.

[32] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Jama*, 270(24):2957–2963, 1993.

[33] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus-Robert Müller. Introduction to machine learning for brain imaging. *Neuroimage*, 56(2):387–399, 2011.

[34] Eleanor A Maguire, David G Gadian, Ingrid S Johnsrude, Catriona D Good, John Ashburner, Richard SJ Frackowiak, and Christopher D Frith. Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, 97(8):4398–4403, 2000.

[35] Manel Martínez-Ramón, Vladimir Koltchinskii, Gregory L Heileman, and Stefan Posse. fMRI pattern classification using neuroanatomically constrained boosting. *Neuroimage*, 31(3):1129–1141, 2006.

[36] Lara Menzies, Samuel R Chamberlain, Angela R Laird, Sarah M Thelen, Barbara J Sahakian, and Ed T Bullmore. Integrating evidence from neuroimaging and neuropsychological studies of obsessive-compulsive disorder: the orbitofronto-striatal model revisited. *Neuroscience & Biobehavioral Reviews*, 32(3):525–549, 2008.

[37] Tom Michael Mitchell. *Machine learning*, volume 1. McGraw Hill, 1997.

[38] John M Ollinger and Jeffrey A Fessler. Positron-emission tomography. 1997.

[39] Graziella Orrù, William Pettersson-Yeo, Andre F Marquand, Giuseppe Sartori, and Andrea Mechelli. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews*, 36(4):1140–1152, 2012.

[40] Emilio Parrado-Hernández, Vanessa Gómez-Verdejo, Manel Martínez-Ramón, John Shawe-Taylor, Pino Alonso, Jesús Pujol, José M Menchón, Narcis Cardoner, and Carles Soriano-Mas. Discovering brain regions relevant to obsessive–compulsive disorder identification through bagging and transduction. *Medical image analysis*, 18(3):435–448, 2014.

[41] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, 45(1):S199–S209, 2009.

[42] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[43] Daniele Radaelli, Alessandro Bernasconi, and Francesco Benedetti. Psychiatric diseases. *Neurological sciences*, 29(3):339–341, 2008.

[44] JH Ranson, KM Rifkind, DF Roses, SD Fink, K Eng, and FC Spencer. Prognostic signs and the role of operative management in acute pancreatitis. *Surgery, gynecology & obstetrics*, 139(1):69–81, 1974.

[45] Jan Sijbers, Paul Scheunders, Noel Bonnet, Dirk Van Dyck, and Erik Raman. Quantification and improvement of the signal-to-noise ratio in a magnetic resonance image acquisition procedure. *Magnetic resonance imaging*, 14(10):1157–1163, 1996.

[46] Le Song, Alex Smola, Arthur Gretton, Karsten M Borgwardt, and Justin Bedo. Supervised feature selection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 823–830. ACM, 2007.

[47] Carles Soriano-Mas, Jesús Pujol, Pino Alonso, Narcís Cardoner, José M Menchón, Ben J Harrison, Joan Deus, Julio Vallejo, and Christian Gaser. Identifying patients with obsessive–compulsive disorder using whole-brain anatomy. *Neuroimage*, 35(3):1028–1037, 2007.

[48] Dan J Stein, Eric Hollander, Chan Stephen, Concetta M DeCaria, Sadek Hilal, Michael R Liebowitz, and Donald F Klein. Computed tomography and neurological soft signs in obsessive-compulsive disorder. *Psychiatry Research: Neuroimaging*, 50(3):143–150, 1993.

[49] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer, 2008.

[50] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[51] Lida Ungar, Paul G Nestor, Margaret A Niznikiewicz, Cynthia G Wible, and Marek Kubicki. Color Stroop and negative priming in schizophrenia: an fMRI study. *Psychiatry Research: Neuroimaging*, 181(1):24–29, 2010.

[52] Vladimir Naumovich Vapnik and Samuel Kotz. *Estimation of dependences based on empirical data*, volume 41. Springer-Verlag New York, 1982.

[53] Prashanthi Vemuri, Jeffrey L Gunter, Matthew L Senjem, Jennifer L Whitwell, Kejal Kantarci, David S Knopman, Bradley F Boeve, Ronald C Petersen, and Clifford R Jack Jr. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *Neuroimage*, 39(3):1186–1197, 2008.

[54] Ze Wang. A hybrid SVM–GLM approach for fMRI data analysis. *Neuroimage*, 46(3):608–615, 2009.