Master in Artificial Intelligence (UPC-URV-UB)

Master of Science Thesis

Automatic Classification of Attention-Deficit/Hyperactivity Disorder using Brain Activation

Maria del Mar Vila Muñoz

Advisor: *Laura Igual Muñoz*

September 4, 2012

Acknowledgements

I am sincerely grateful to my supervisor, Laura Igual, whose support and orientation from the beginning until the end allowed me to understand and develop this project, and also for her patience and friendly advices.

I also would like to thank Petia Radeva who got me into the exciting field of Artificial intelligence and Computational Neuroscience. Thank you both for giving me some of your valuable time.

Finally, dedicate this work to my family, friends and all those who have supported me in some way or another during the project.

Abstract

Nowadays, there is an active field of research in neuroscience trying to find relations between neurofunctional abnormalities of brain structures and neurological disorders. Previous statistical studies on brain functional Magnetic Resonance Images (fMRI) have found Attention Defficit Hyperactivity Disorder (ADHD) patients are characterized by reduced activity in the inferior frontal gyrus (IFG) during response inhibition tasks and in the Ventral Striatum (VStr) during reward anticipation tasks.

Interpreting brain image experiments using fMRI requires analysis of complex data and different univariate or multivariate approaches can be chosen. Recently, one analysis approach that has grown in popularity is the use of machine learning algorithms to train classifiers to discriminate abnormal behavior or other variables of interest from fMRI data.

The purpose of this work is to apply machine learning techniques to perform fMRI group analysis in an adult population. We propose a multivariate classifier using different discriminative features. Furthermore, we show how temporal information of fMRI data can be taken into account to improve the discrimination.

We demonstrate that our new approach is able to diagnose the ADHD characteristics based on the activation in the executive functions. Previous results on the response inhibition task did not find differences between activation responses. Opposite to these results, we achieve accurate classification performance for this task. Moreover, in this case, we show that classification rates can be significantly improved by incorporating temporal information into the classifier.

Contents

Chapter 1

Introduction

Attention-deficit/hyperactivity disorder (ADHD) is a development disorder characterized by inattentiveness, motor hyperactivity and impulsiveness and it represents the most prevalent childhood psychiatric disorders. It is estimated that in a 50% of the cases its symptoms persists into adulthood. These symptoms come from disruptions in executive functions, especially poor inhibition control and abnormalities in the motivational system related to reward anticipation, each one developed in a Region Of Interest (ROI). The first, developed in Inferior Frontal Gyrus (IFG) and the second in Ventral Striatum (VStr). Both functions, response inhibition and reward anticipation, can be addressed in a subject with the specifics tasks which called Go-NoGo task and Monetary Incentive Delay (MID) task as done in [1].

Functional Magnetic Resonance Imaging (fMRI) is a non invasive technique based in the Blood Oxygenation Level Dependent (BOLD) effect [2]. It has enabled scientists to look into the active human brain by providing sequences of 3D brain images. This has revealed exciting insights into the spatial and temporal changes underlying a broad range of brain functions, such executive functions or brain states.

A wide range of statistical methods are being increasingly applied to the analysis of fMRI time series. For example, Statistical Parametric Mapping (SPM)[15] is a tool based on the General Linear Model (GLM) [3] which can fit the response signal from a region of the brain given an experiment design [3]. These methodological developments are providing cognitive neuroscientists with the opportunity of tackling new research questions that are relevant for the understanding of the functional organization of the brain. fMRI data were originally analysed in statistics and machine learning. In practice,the analysis depend on how well the classifier's implicit assumptions hold in the domain in which is applied and on how much data are available [4].

In the literature about machine learning framework most of the approaches for fMRI data analysis are multivariate classifiers based on dimensionality reduction, on feature selection techniques and on the signal of the region without using GLM. These studies are always related to detect brain states in a subject when he receive a specific stimulus. Some approach is made using the GLM obtained from a experiment design as in [5], but in this case they do not use temporal information and also is for brain state applications. To our knowledge, [6], which is a recent study, is the first study that learning on fMRI data that is used for classification purposes. They perform a group analysis, separating drug addicted subjects from healthy non-drug-using controls, and it is performed explicitly on temporal information observed fMRI time sequences in a ROI.

Regarding SPM, once the subject has been exposed to the experimental task and the fMRI data is acquired, it performs a statistical analysis about the signal in a ROI. It fits the ROI to an activation pattern (activation-rest-activation-...) based, as we said, on GLM. In particular, given an experiment design it return as an output all the parameters from the GLM. These are, for example the beta parameters, which are the parameters from the regressors in the BOLD signal. Most of beta parameters, each reflecting the activation level (effect size) of the ROI, come from specific conditions of the experiment (for example when a subject is waiting a signal to do something or when he gains a reward). Then the contrasts, also obtained by SPM, allow analysing the differences between two (or more) activation patterns defined by betas (for example, which is the effect size with a specific condition respect to the rest condition). Finally, fitted data gives the activation pattern for a ROI modelled by GLM with specific contrast.

In addition to analyse the BOLD signal in a single subject, SPM also is able to perform a subject group comparison (second level analysis). In [1], authors use SPM to see, first if there is activity (i.e. the activation pattern is significant) in each ROI (first level analysis) and second, to perform a second level analysis to find significant differences in BOLD signal between ADHD patients (adult medical-naive patients) and controls (group analysis). As a result, they find that activation pattern in both ROIs, IFG and VStr, during the tasks; there is a reduced VStr activity in the ADHD group; and they did not find significant differences between groups in IFG activity.

Our purpose is to apply machine learning techniques to perform the group analysis using data in [1]. We propose a multivariate classifier based on the GLM parameters obtained in the first level analysis. We use different features: the beta parameters, the contrasts, and also temporal features such as fitted data. Thereby, we want to show if using a machine learning approach and analysing the temporal changes we can obtain an accurate classifier between the two groups, ADHD patients and controls. Thus, we want to validate the hypothesis of [1] for VStr activity and IFG activity.

Chapter 2

Problem definition

2.1 ADHD and Functionality of ADHD

What is ADHD?
Attention-deficit/hyperactivity disorder (ADHD) is a syndrome characterized by symptoms of inattention, hyperactivity, and impulsivity. This syndrome is one of the most prevalent childhood psychiatric disorders. Around 5% of children are diagnosed worldwide [7].

Although symptoms commonly tend to improve with age, only a minority of ADHD children attain to complete remission in adult life [8]. In fact, a recent 10-year follow-up study indicates that symptoms persist into adulthood in more than 50 % of cases, and reach complete remission around adult life.

Given the disabling nature of ADHD, it is important to understand the neural base of the disorder, particularly in those subjects that do not ameliorate with age.

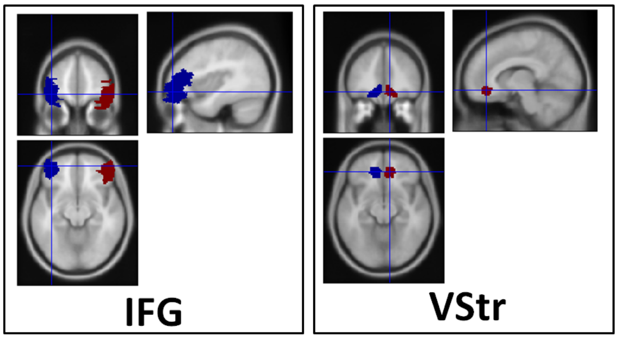


Figure 2.1: At Left, left and right Inferior Frontal Gyrus. At Right, Ventral Striatum VStr. Bothe images are obtained using Marsbar tool from Matlab and correspond to one of the subjects for the experiment.

Biology and analysis

Based mostly on studies of children with ADHD, current accounts of the disorder propose the implication of at least two relatively independent but not mutually exclusive ADHD endophenotypes[1]: those characterized by disruptions in executive functions, especially poor inhibition control, and those characterized by abnormalities in the motivational system, particularly in relation to reward anticipation.

Previous functional magnetic resonance imaging (fMRI) studies of ADHD have predominantly focused on the assessment of executive functions, in particular response inhibition, which is often addressed using the *Go-NoGo* task. This paradigm has been shown to rely on *inferior frontal functioning* (Figure 2.1) in the healthy population, and ADHD imaging studies have consistently found hypoactivation in this region during childhood phase of the disorder [9].

On the other hand, the *Monetary Incentive Delay* (MID) task is used in ADHD research to investigate the neural bases of reward management. It is used to target reward-related regions such as the *Ventral Striatum* (VStr) (Figure 2.1) and it has been used by different groups to study motivational processes in ADHD. To date, the few studies assessing reward anticipation in ADHD have reported reduced recruitment of the VStr in both child and adult patients as compared to control subjects, and nearly all such studies have observed a negative association between VStr activation and hyperactive/impulsive symptoms [10, 11, 12].

Two new paradigms has emerged related to the ADHD studies which are also in [1]. The first is that both the Go-NoGo and MID task have previously been applied in ADHD research, but no studies have implemented those tasks in an *intra-subject* manner. The intra-subject application of the two tasks is important to exclude the variability between subjects and between studies. Secondly no previous fMRI studies have applied either of these tasks in adults with ADHD who have never received medication for their condition. Since methylphenidate/atomoxetine administration has been shown to render short-term and long-term synaptic, structural, and functional changes in key region like inferior frontal gyrus (IFG) [13, 14] and VStr , the assessment of medication-naïve patients is essential to relate IFG and VStr alterations to the neurobiology of the disorder.

In [1], the paradigms cited above are addressed. They explore weather medication-naïve adults with ADHD exhibit behavioural and neural distur-

[1] Endophenotype is a genetic epidemiology term which is used to parse behavioral symptoms into more stable phenotypes with a clear genetic connection. A phenotype is the composite of an organism's observable characteristics or traits: such as its morphology, development, biochemical or physiological properties or behavior.

bances in both response inhibition and reward anticipation. Their hypothesis is that unmedicated adults with ADHD show deficient IFG activation during response inhibition and deficient VStr activation during reward anticipation. However in the results, they only observe reduced bilateral VStr activity in adults with ADHD during reward anticipation and no differences are detected in IFG activation during response inhibition task.

In the present study, we propose to use machine learning techniques in the same fMRI data and the same paradigms used in [1]. Our aim is to apply these new techniques to their result on reward anticipation and check the hypothesis that unmedicated adults with ADHD will show deficient IFG activation during response inhibition.

2.2 Medical Imaging Techniques

2.2.1 MRI

Magnetic resonance imaging (MRI) is a no invasive technique which allows to study the body structure under different parameters and any spatial orientation.

As it is stated in [2], MRI uses nuclear magnetic resonance (NMR) signals to create images of the brain and uses hydrogen nuclei as the basis for the signal.

Elementary particles (electrons, protons and neutrons) have a quantum property called *spin*. For a particular atom, these particles combine each other in pairs of opposite spin. The hydrogen nuclei's spin make them to behave like small magnetic dipoles which is the basis for the NMR signal.

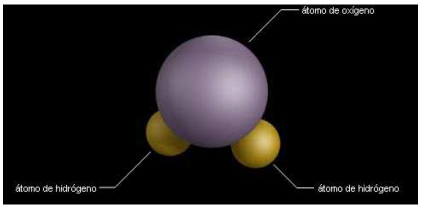


Figure 2.2: Water molecule

Hydrogen is chosen because is the most abundant nuclei with that spin in body tissue due to the water. Humans composed by a between 60 and 70 percent water and it is formed by a particle of hydrogen and two oxygen as is shown in Figure 2.2

Note that the following is a classical physics description of a phenomenon that can only be accurately described by quantum physics. While we cannot really know how individual protons are behaving, this is an approximation of the net action of a lot of protons, just useful for visualization.

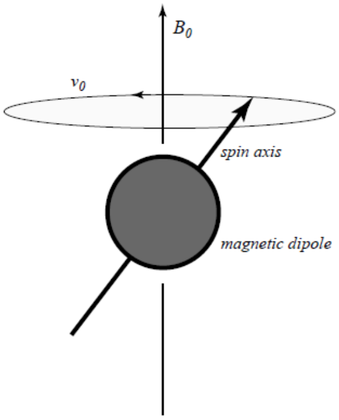


Figure 2.3: The magnetic field B_0 exerts a torque on a nuclear magnetic dipole that would tend to make it align with B_0. However, because the nucleus also has angular momentum (*spin*), it instead precesses like a spinning top at an angle to the gravitational field. The precession frequency v_0 is proportional to the magnetic field and is the resonant frequency of NMR. [2]

In MRI, the brain is placed in string magnetic field B. Hydrogen nuclei (like small magnets) align with the magnetic field producing their own net *longitudinal* magnetization (in the same direction as B). This alignment, a relaxation towards the equilibrium state, occurs with a time constant T1. Full alignment is never reached and the nuclei precess around the axis of the field at a frequency (v) proportional to the strength of the magnetic field (Figure 2.3). The precession frequency (v), also called Larmor frequency, is given by:

$$v = \gamma B \tag{2.1}$$

where γ is a constant (the gyromagnetic ratio equal to 42.6Mz/T for the hydrogen nucleus). v = 128 MHz in a 3 Tesla[2] field.

[2] The Tesla (symbol T) is the International System of Units (SI) derived unit of magnetic flux density.

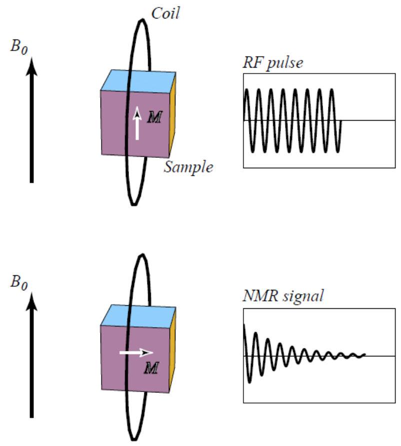


Figure 2.4: A sample is placed in a large magnetic field B_0, and hydrogen nuclei partially align with the field creating a net magnetization M. In the transmit part of the experiment an oscillating current in a nearby coil creates an oscillating RF magnetic field in the sample which causes M to tip over and precess around B_0. In the receive part of the experiment, the precessing magnetization creates a transient oscillating current (the NMR signal) in the coil

Next in MRI, a radio frequency pulse (generated by an oscillating current in a coil) is transmitted to the sample at the resonant frequency. The Radio Frequency (RF) pulse tips over the acquired longitudinal magnetization by 90 degrees into the transverse plane. The magnetization now precesses around B in the transverse plane (Figure 2.4). The precessing nuclei absorb that energy and re-emit a portion of it back at the same frequency, this is the detected signal. Due to this precessing transverse magnetization, a detector located in the transverse plane will feel a small oscillating magnetic field. The changing magnetic field induces a current in the detector coil (electromagnetic induction) that can be measured.

The resulting MR image will be an image of the transverse magnetization of hydrogen nuclei at the time when the signal is detected. The transverse magnetization is a transient phenomenon and now we will see how the strength of the signal depends critically on timing parameters during imaging. [2]

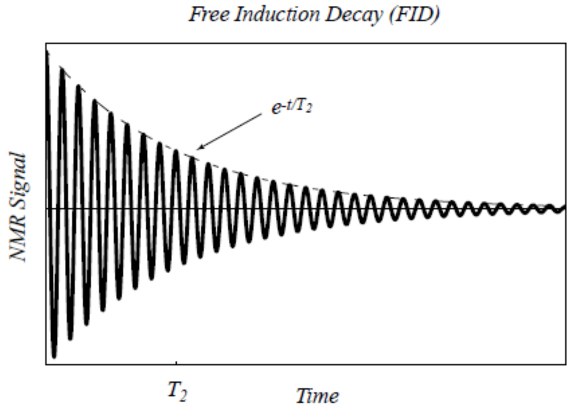


Figure 2.5: The FID. After a 90° RF pulse tips the longitudinal magnetization into the transverse plane, a detector coil measures an oscillating signal which decays in amplitude with a time constant T2 in a perfectly homogeneous magnetic field. The plot is not to scale; typically the signal will oscillate more than a million times during the interval T2. [2]

After the magnetization is tipped by the RF pulse, the transverse component (and thus the signal) decays with a time constant T2 and the longitudinal component regrows (towards equilibrium) with a time constant T1. The signal decay is called *Free Induction Decay* (FID) and a plot example is shown in Figure 2.5.

As shown in Figure 2.6, in a MR imaging sequence successive RF pulses are transmitted and successive FID signals are measured. The strength of the signal measured depends critically on the imaging parameters TE (time between RF pulse and measurement) and TR (time between successive RF pulses).

The time constants T2 and T1 are on the order of one second but the actual rates depend on intrinsic properties of the tissue surrounding the nuclei. As a result, the strength of the signal will vary for different body tissues with different intrinsic T2 and T1 values (i.e. gray matter compared to white matter) creating the primary contrast in an MR image of the brain. Figure 2.7 shows MR images of the same anatomical section showing a range of tissue contrasts varying T1, T2 and also TR and TE. When the contrast can be manipulated by varying TR and TE:

- If TR is short, the longitudinal magnetization will not have a chance to fully regrow between pulses. The size of the next FID signal will be reduced (because there is less magnetization to tip) by an amount that depends on the T1 of each tissue (see Figure 4). If TE is also short, little T2 decay will have occurred before measurement. Thus, the contrast in the MR image will depend primarily on the intrinsic T1 values of the tissues. Such an image is said to be *T1-weighted.*

- If TR is long, the longitudinal magnetization will fully regrow between pulses and T1 will provide no image contrast. If TE is long enough for some (but not all) T2 decay occur, then the signal strength will depend on the intrinsic T2 values of the tissues. The resulting image is said to be *T2-weighted.*

- If TR is long, so that the magnetization fully regrows between pulses, and TR is short, so that Little T2 decay occurs before measurement, then the signal will depend little on the T1 and T2 values of the tissue. Rather it will depend absolute strength of the acquired magnetization in each tissue (which depends primarily on proton density). The resulting image is said to be *density weighted.*

MRI is not based on a single parameter as the attenuation coefficient of X-ray, but some independent parameters like TI, T2, TR and TE vary considerably from one tissue to another in MRI. While the absorption coefficient of X-rays varies only 1% between different tissues, the spin density and relaxation time T1 of these tissues differ by 20-30%. The T2 relaxation time differs by 40% for the same tissues. These differences are responsible for its excellent low-contrast resolution which is the main advantage of this technique.

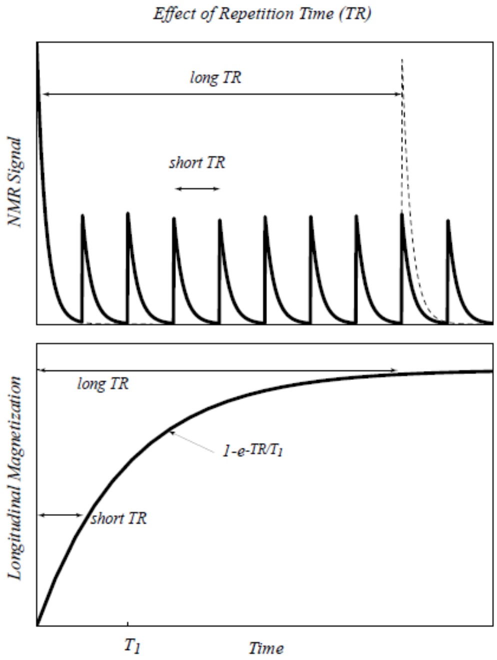


Figure 2.6: The effect of *Time Repetition* (TR). Repeated RF pulses generate repeated FID signals, but if the TR is short, each repeated signal will be weaker than the first (top). The magnitude of the signal with a 90° RF pulse, is proportional to the magnitude of the longitudinal magnetization just prior to the RF pulse. After a 90° RF pulse the longitudinal magnetization recovers toward equilibrium with a relaxation time T1(bottom). If this recovery is incomplete because TR<T1, the next FID signal is reduced. [2]

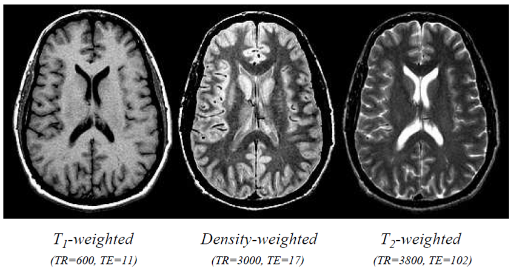


Figure 2.7: In the first image (left to right) Cerebrospinal Fluid (CSF) is black, while in the last image CSF is bright. Contrast is manipulated by adjusting several parameters during image acquisition, such as the repetition time TR and the echo time TE (times given in milliseconds), which control the sensitivity of the signal to the local tissue relaxation times T1 and T2, and the local proton density. [2]

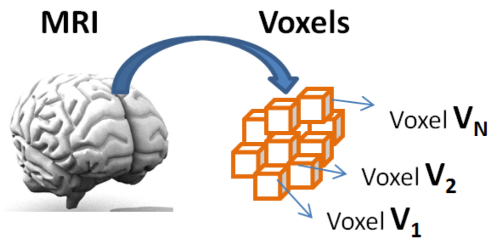


Figure 2.8: Here there is a visual representation of voxels from a MRI.

In the way to represent a MRI, are used small geometric elments called *voxels*. A *voxel* is a volumetric pixel representing a value on a regular grid three dimensional space and it is used for the visualization and analysis of medical and scientific data. MRI is a 3D image and formed by a lot of small cubes as is shown in 2.8.

2.2.2 Functional MRI

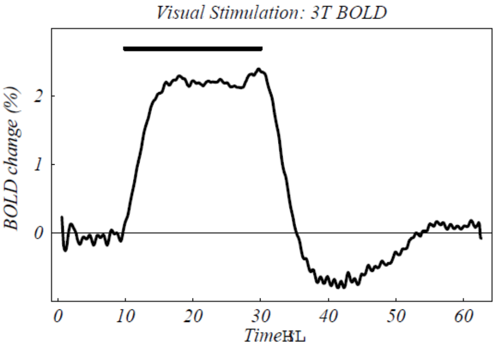


Figure 2.9: Example BOLD response in the visual cortex measured at 3T. Subjects wore goggles that flashed a grid of red lights at 8 Hz. The stimulus (indicated by a horizontal bar) lasted for 20 sec, followed by 40 sec of darkness. The data shows the average response of 32 cycles of stimulus/rest for 3 subjects. Characteristic features of the BOLD response are a delay of a few seconds after the start of the stimulus, a ramp of about 6 sec up to a plateau, and a post-stimulus undershoot before the signal returns to baseline

Functional MRI (fMRI) is a non-invasive technology born in 1990 which goal is to understand the complex relationships between neuronal activity, metabolism, and blood flow. Until now, the accumulating evidence suggests that the blood oxygenation response is roughly a linear function of neuronal activity (at least to a first approximation).

The basic idea behind fMRI is simple: we measure a series of MRIs (like a movie) and look for small changes in MR signal intensity over time caused by changes in brain activity. This change in the MR signal depends on two factors:

- First, brain activity is accompanied by a local increase in blood flow that supplies oxygen at a rate faster than its consumption. As a result of the oversupply, brain activity leads to an increase in the local concentration of oxygenated haemoglobin compared to de-oxygenated haemoglobin in venous blood.
- Second, deoxygenated hemoglobin is paramagnetic and attenuates the MR signal more quickly than oxygenated hemoglobin (because it generates local field inhomogeneities). As a result, the increase in the ratio of oxygenated/deoxygenated hemoglobin leads to a local increase in the

MR signal (the image becomes brighter). This change in signal intensity is called the *Blood Oxygenation Level Dependent* (BOLD) effect. An example of BOLD effect is shown in Figure 2.9

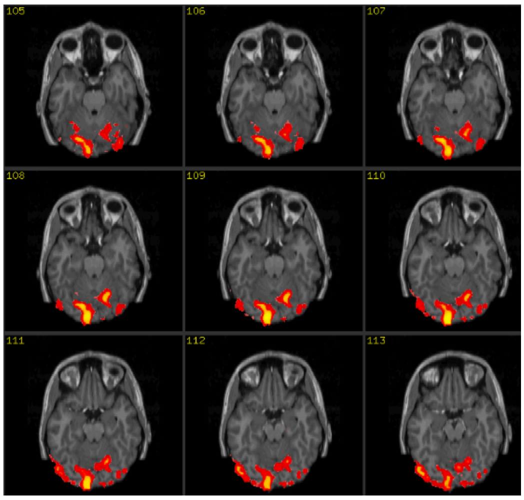


Figure 2.10: An axial plane of fMRI in different time points.

fMRI as MRI is represented by *voxels* and fMRI is a set of MRIs (in consecutive different time points) measuring intensity changes in a same voxel across the time. Thus MRI studies brain anatomy while fMRI studies brain function. A single voxel's response signal over time is called its *timecourse*. In Figure 2.10 there is an example of intensity changes from a group of voxels.

Chapter 3

State of the Art

The analysis of brain activation experiments by means of fMRI data was originally faced using statistics studies. Different useful softwares, as the powerful Statistical Parametric Mapping (SPM) [15], have helped in this work carried out mainly by phycologist and medical doctors, which are used to work in the statistical field.

Different univariate analysis approaches have been presented to analyse several mental disorders [16]. Functional MRI has been a standard tool for visualizing regional brain activations during sensorimotor or cognitive tasks. The properties of fMRI data are well understood and, usually, modeled by using General Linear Model (GLM) [17, 18, 19]. GLM based approach consists on the following: fMRI time series data are fitted to an a priori defined reference function; the fitting parameters are then contrasted to produce a test statistic at each voxel. In [20], fMRI data are widely processed using a general linear model (GLM) based approach. Assuming linear brain responses to the functional stimuli and modelling each voxel's time series with a canonical HRF [21, 22], this univariate GLM approach may by suboptimal for fMRI data analysis. To consider the HRF discrepancies, a HRF model-free approach is highly preferred for fMRI data analysis.

Recently, machine learning algorithms have grown in popularity to analyse behaviours and other variables of interest from fMRI data. However, performance of machine learning techniques, in practice, will depend on how well the classifier's implicit assumptions hold in the domain in which is applied and on how much data are available [4].

Following that, [23] states that fMRI data analysis methods can be roughly divided into two main classes: the hypothesis-driven methods and the exploratory methods. The hypothesis-driven fMRI data analysis methods, represented by the conventional GLM, have a strictly defined statistical framework for assessing regionally specific activations but require prior brain response modeling that is usually hard to be accurate [24]. On the contrary, exploratory methods, like the support vector machine (SVM), are independent of prior hemodynamic response function (HRF). In [20], the hybrid SVM-GLM combines both to take the advantages of both kinds of methods. SVM is used to learn the data-driven response function and GLM is applied to fitted the signal. SVM-GLM showed better results than regular GLM for detecting the sensorimotor activations. Moreover, SVM is a machine learning-based auto-classification method which has been demonstrated to be useful for analysing neuroimaging data in many applications. Particularly, it has shown good promise for exploring the spatial brain discriminance patterns (SDP) between different populations or between different brain states [25, 20, 26, 27, 6, 28, 29, 30].

On the other hand, the fMRI data are multivariate in nature since each fMRI volume contains information about brain activation at thousands of measured locations (voxels). Multivariate techniques have been applied to neuroimage data in many studies (for review, see [31]). For example, [18] introduced a multivariate approach using standard multivariate statistic and GLM to make inferences about effects of interest and canonical variates analysis (CVA) to describe the important feature of these effects. [32] describes a multivariate method for analyzing fMRI data based on independent components analysis (ICA) [33], which can be used to distinguish between non-task-related signal components, movements and other artifacts, as well as consistently or transiently task-related fMRI activations.

The previous use of classifiers for fMRI data analysis can be divided in two groups. The first group applied classifiers after preprocessing using a feature selection methods based on prior hypotheses [27, 34, 25, 35]. Also in [36, 37, 38, 39] a multivariate feature selection has also been applied to fMRI data. In these studies, the data were encoded as a vector of features, one feature for each voxel (hundred of thousands of features). Because of the high dimensionality of this feature vector, a feature selection method based on prior hypotheses was applied to the data set to reduce the dimensionality, and afterwards the selected features were used as inputs to a classifier, that is, the discriminating regions were chosen a priori and given as input to the classifiers. The second group used Principal Components Analysis (PCA) or Singular Value Decomposition (SVD) analysis as dimensionality reduction method and applied the classifier on PCA/SVD basis without prior selection of spatial features, (e.g. [40, 41, 42, 43]). They introduced the concept of models of functional activation for classification. In [30], authors use a SVM algorithm to perform multivariate classification of brain states from whole fMRI volumes without prior selection of spatial features. This classifier predicts the subject's instantaneous brain state.

In [5]], the response pattern from stimuli with multivariate classifier is decoded in specified ROI using beta values. They compare 6 multivariate classifiers and investigate how response how response-amplitude estimated (beta-value) affect classification performance. The classifiers are: pattern correlation classifier, k-nerest neighbours, Fisher's linear discirminant, Gaussian naïve Bayes, and linear and non-linear (radia-basis-function kernel) SVM. Each method is evaluated for different combinations of several variables (multivariate) as: ROI, pattern normalization, cross-validation scheme, categorical stimulus. Also in the literature, [44, 37], some pattern-information studies used beta-estimates to define the response patterns

In [45], there is an analytical strategy focused on the mapping of a "stimulus-single location response". A statistical pattern recognition exploits and integrates the information available at many spatial locations thus allowing the detection of perceptual and cognitive differences that may produce only weak single-voxel effects. Analysing the relation between a stimulus and the responses simultaneously measured at many locations (spatial response patterns or multivolxel response patterns) allow to localize effects that may remain invisible to the conventional analysis with univariate statistical methods. This approach is named *Multivariate Pattern Analysis* (MVPA). In this approach, a spatially invariant model of the BOLD response is fitted independently at each voxel's time course (massively univariate analysis), and using machine learning and pattern recognition techniques (SVM) differences between estimated activation levels during two or more experimental conditions are tested.

From the point of view of group analysis, to our knowledge, [6] is the first approach in which learning on fMRI data is performed explicitly on temporal information for classification. The application is group analysis: separating drug addicted subjects from healthy non-drug-using controls based on their observed fMRI time sequences. By selecting discriminative features, group classification can be successfully performed on the case of study, although training data are exceptionally high dimensional, sparse and noisy fMRI sequences.

Chapter 4

Methodology

4.1 Statistical Parametric Mapping

Inferences in neuroimaging may be about differences expressed when comparing one group of subjects to another or, within subjects (intra-subject), changes over a sequence of observations. They may pertain to structural differences (e.g. in voxel-based morphometry) or neurophysiological measures of brain functions (e.g. fMRI).

As it is stated in [3] Statistical parametric mapping (SPM [1]) is a parametric map of physiological or physical parameters (e.g. parametric maps of regional Cerebral Blood Flow (rCBF) or volume). SPM is used to identify regionally specific effects in neuroimaging data and is a prevalent approach for characterizing functional anatomy, specialization and disease-related changes. Specifically SPM is a voxel-based approach, employing topological inference, to analyse regionally specific responses to experimental factors. In order to assign an observed response to a particular brain structure, or cortical area, the data are usually mapped into an anatomical space.

Functional mapping studies are usually analysed with some form of SPM[2]. Thus, SPM entails the construction of continuous statistical processes to test hypotheses about regionally specific effects. SPMs are images (Figure 4.1) or fields with values that are, under a null hypothesis, distributed according to a known probability density function, usually the Student's t or F-distributions. These are known colloquially as *t-* or *F-maps.*

[1] *SPM* is also referred to the software package from Matlab and it is used to compute statistical parametric map

[2] Parametric statistics is a branch of statistics that assumes that the data has come from a type of probability distribution and makes inferences about the parameters of the distribution.

Over the years, statistical parametric mapping has come to refer to the conjoint use of the *General Linear Model* (GLM) and *Random Field Theory* (RFT)[3].

This provides analysis and make classical inferences about topological features of the SPM. The GLM is used to estimate some parameters that explain continuous data in exactly the same way as in conventional analyses of discrete data. RFT is used to resolve the multiple comparison problem making inferences over the volume analysed.

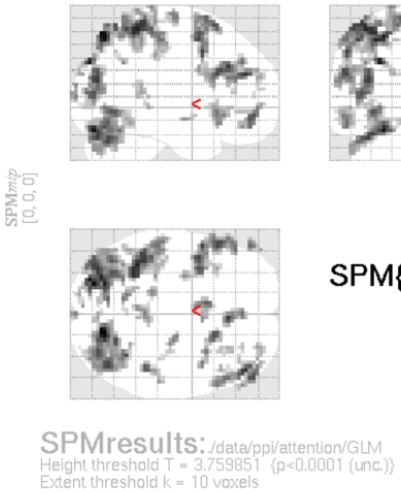

SPMmip
[0, 0, 0]
SPM{T341}
SPMresults: ./data/ppi/attention/GLM
Height threshold T = 3.759851 {p<0.0001 (unc.)}
Extent threshold k = 10 voxels

Figure 4.1: Statistical Parametric Map extracted from SPM interface which represents the region of the brain which is activated as experimental task response. [15]

4.2 GLM and SPM Analysis

The GLM is a strategy to express a measure (such as rCBF) in a experiment conducted. The GLM is based in the equation

$$Y = XB + \varepsilon \tag{4.1}$$

[3] A random field is a generalization of a stochastic process which is a collection of random variables. Instead of describing a process which can only evolve in one way (as in the case, for example, of solutions of an ordinary differential equation), in a stochastic or random process there is some indeterminacy

that expresses the observed response variable in terms of a linear combination of *explanatory variables* X plus a well behaved error term ε.

The matrix X that contains the *explanatory variables* (e.g. designed effects or confounds) is called the *design matrix*. Each column of the *design matrix* corresponds to an effect that build the experiment or that may confound the results. These are referred to as explanatory variables or *regressors*. The relative contribution of each of these columns is assessed using standard maximum likelihood and inferences about these contributions using t or F-statistics, depending upon whether one is looking at a particular linear combination, or all of them together.

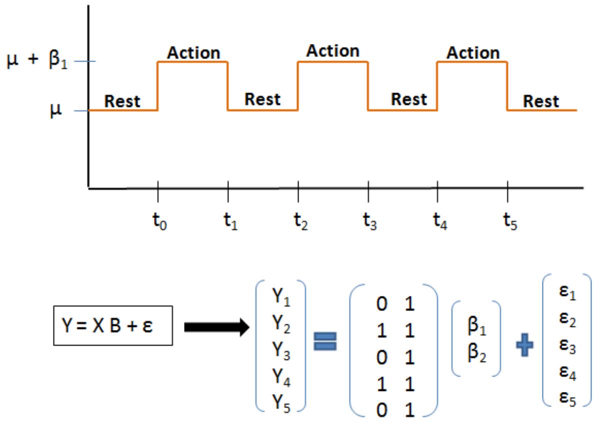


Figure 4.2: It is a simple case (univariate approach, i.e. for only one voxel) where GLM represents an experimental task with *action* and *rest* condition. In This case there are only 2 regressors. In his case if we want to test the activation pattern, we need to compare the condition *action* and condition *rest*, so the contrast used would be $[1, -1]$.

Each column l from the design matrix X has associated a parameter which is the lth element from B. Each element from B is called β. Some of these parameters will be of interest (e.g. the effect of a particular sensorimotor or cognitive condition or the regression coefficient of haemodynamic responses on reaction time). This means that any condition effect of interest can be represented using its correspondent β (Figure 4.2 shows the simplest case where the design matrix represents an experimental task). Each parameter of interest also reflects the activation level (in terms of rCBF) for a specific condition, this is also called the *effect size* for this condition. The remaining parameters will be of no interest and pertain to *confounding effects* (e.g. the effect of being a particular subject or the regression slope of voxel activation on global activity), these parameters are removed by SPM and are not shown in the design matrix used to do the analysis.

Is needed to keep in mind that it's an experiment to measure a response variable which is the rCBF at a particular voxel Y_j where $j = 1, \ldots, J$ indexes the observation, and Y_j is the jth element from the vector Y. Suppose now that, for each observation, we have a set of L ($L < J$) explanatory variables (each measured without error) denoted by x_{jl}.
Then GLM explains the response variable Y_j in terms of linear combination of the explanatory variables plus an error term as follows:

$$Y_j = x_{j1}\beta_1 + \ldots + x_{jl}\beta_l + \ldots + x_{jL}\beta_L + \epsilon_j \tag{4.2}$$

Here the β_l, $l = 1, \ldots, L$, are (unknown) parameters, corresponding to each of the L explanatory variables x_{jl}. The errors ϵ_j are independent and identically normal distributed with zero mean and variance σ^2, written $\epsilon_j \overset{idd}{\sim} N(0, \sigma^2)$.

The design matrix X defines the experimental design and the nature of the hypothesis testing and Y is the measured response. The goal is to estimate β parameters.

4.2.1 Parameter Estimation

Usually, the simultaneous equations implied by GLM (with $\epsilon = 0$) cannot be solved, because the number of parameter L is typically less than the number of observations$J$. Therefore, some method of estimating parameters that 'best fit' the data is to required. This can be achieved by the method of *ordinary least squares*.

Denote the set of parameters estimates by $\tilde{\beta} = [\tilde{\beta}_1, \ldots, \tilde{\beta}_L]^T$. Those parameter lead to *fitted values* $\tilde{y} = [\tilde{Y}_1, \ldots, \tilde{Y}_J]^T = X\tilde{\beta}$, where X is the design matrix and giving the residual errors $\epsilon = [\epsilon_1, \ldots, \epsilon_J]^T = Y - \tilde{Y} = Y - X\tilde{\beta}$. The *residual sum-of-squares* $S = \sum_{j=1}^{J} \epsilon_j^2 = \epsilon^T\epsilon$ is the sum of squares differences between the observed and fitted values[4]. The *least squares* estimates are the parameter estimates which minimize the residual sum-of-squares, i.e.:

$$S = \sum_{j=1}^{J} (Y_j - x_{j1}\tilde{\beta}_1 - \ldots - x_{jL}\tilde{\beta}_L)^2 \tag{4.3}$$

[4] $\epsilon^T\epsilon$ is the L_2 norm which is equivalent to the distance between the model and the data

And this is minimized when:

$$\frac{\partial S}{\partial \tilde{\beta}_l} = 2\sum_{j=1}^{J}(-x_{jl})(Y_j - x_{j1}\tilde{\beta}_1 - \ldots - x_{jL}\tilde{\beta}_L) = 0 \tag{4.4}$$

This equation is the l^{th} row of $X^TY = (X^TX)\tilde{\beta}$. Thus, the least squares estimates, denoted by $\tilde{\beta}$, satisfy:

$$X^TY = (X^TX)\tilde{\beta} \tag{4.5}$$

For the GLM, the least squares estimates are the *maximum likelihood estimates.* If (X^TX) is invertible, which it is if, and oly if, the design matrix X is full rank, then the least squares estimates are:

$$\tilde{\beta} = (X^TX)^{-1}X^TY \tag{4.6}$$

4.2.2 Contrasts

For an independent and identical error, the residual variance σ^2 is estimated by the residual sum-of-squares divided by the appropriate degrees of freedom: $\tilde{\sigma^2} = \frac{e^Te}{J-p} \sim \sigma^2\frac{X^2_{J-p}}{J-p}$ where $p = rank(X)$ and X the design matrix.

It is not too difficult to show that the parameter estimates are normally distributed: if X is full rank then $\tilde{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$. From this it follows that for a column vector c, named *contrast vector* containing L weights:

$$c^T\tilde{\beta} \sim N(c^T\beta, \sigma^2c^T(X^TX)^{-1}c) \tag{4.7}$$

Furthermore, $\tilde{\beta}$ and $\tilde{\sigma}^2$ are independent (Fisher's law). Thus, prespecified hypothesis concerning linear combination of the model parameters $c^t\beta$ can be assessed using:

$$\frac{c^T\tilde{\beta} - c^T\beta}{\sqrt{\tilde{\sigma}^2c^T(X^TX)^{-1}c}} \sim t_{J-p} \tag{4.8}$$

where t_{J-p} is a Student's t-distribution with $J-p$ degrees of freedom. For example, the hypothesis $H : c^T\beta = d$ can be assessed by computing

$$T = \frac{c^T\tilde{\beta} - d}{\sqrt{\tilde{\sigma}^2 c^T (X^T X)^{-1} c}} \tag{4.9}$$

and computing a p-value by comparing T with a t-distribution having $J-p$ degrees of freedom. In SPM, all null hypotheses are of the form $c^t\beta = 0$.

This allows one to test the null hypothesis, that all the estimates are zero when is needed to show the significance of activation pattern for a specific condition. The t-statistic uses a *contrast* (a linear combination expressed as a vector specifying the contrast weights). These contrasts allow to compare the difference in responses between some conditions in the experimental task, or to see if the activation the significant for only one condition.

While beta parameters of interest (section 4.2) from the GLM show the effect size (activation level) for a specific condition, the contrast is able to compare the activation level from two or more beta parameters of interest. An example of a contrast weight vector would be $[1, -1, 0, \ldots, 0]$ to compare the difference in responses evoked by two conditions, modelled by the firsts two condition-specific regressors in the design matrix (defined by β_1 and β_2). Sometimes is needed to analyse if the effect size in a voxel is significant for a specific condition, which means if there is activation for that condition in that voxel, in this case the contrast is $[0, \ldots, 1_l, \ldots, 0]$, where lth is the regressor corresponding to that condition.

4.2.3 Convolution, Noise and Filtering

A timecourse (section 2.2.2) defined by a voxel signal typically has some unwanted signal called *noise*, produced by the scanner, random brain activity and similar elements, which are big as the signal itself (Figure 4.3). These are removed from fMRI data when is neeeded to analysed only the data included in the experimental design.

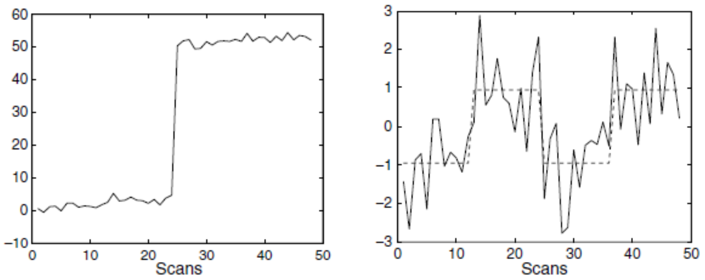


Figure 4.3: Example to show the noise in a signal. Both figures represent the same timecourse signal for a voxel. In the left the exactly signal extracted from the intensity of the voxel during timeis presented and in the right is the signal removed the noise (the dashed line is the fitted data and the solid line, the adjusted data), in this case we can appreciate a clear activation pattern of the signal. [3]

fMRI time-series can be viewed as a linear mixture of signal and noise (as explained before the signal would correspond to the parameters of interest, and the noise the remaining parameters which are not shown in the design matrix). Since data is a mixture of activation and noise that share some frequency bands a *filtering* is needed in order to increase the sensitivity. In the following paragraphs we tell the basis of this filter.

Signal corresponds to neuronally mediated haemodynamic changes that can be modelled as a convolution of some underlying neuronal process, responding to changes in experimental factors, by a haemodynamic response function.

There are two important considerations that arise from this signal perspective on fMRI time-series: the first pertains to optimal experimental design and the second to optimum de-convolution of the time-series to obtain the most efficient parameter estimates.

After the stimulus functions have been specified in terms of onsets and durations, we need to describe the shape of the expected response. This is done using temporal basis functions. The underlying assumption is that the BOLD response follows the haemodynamic[5] response function. The basis function is used to convert the assumed neural activity into haemodynamic activity and this is expressed as regressors in the design matrix. The optimum design in fMRI should present those signals that survive (pass the filter) convolution with the *haemodynamic response function* (HRF). Figure: 4.4 shows two GLMs; the first one is the simplest one and the second is a more

[5] Haemodynamics is a medical term for the dynamic regulation of the blood flow in the brain. It is the principle on which functional magnetic resonance imaging is based

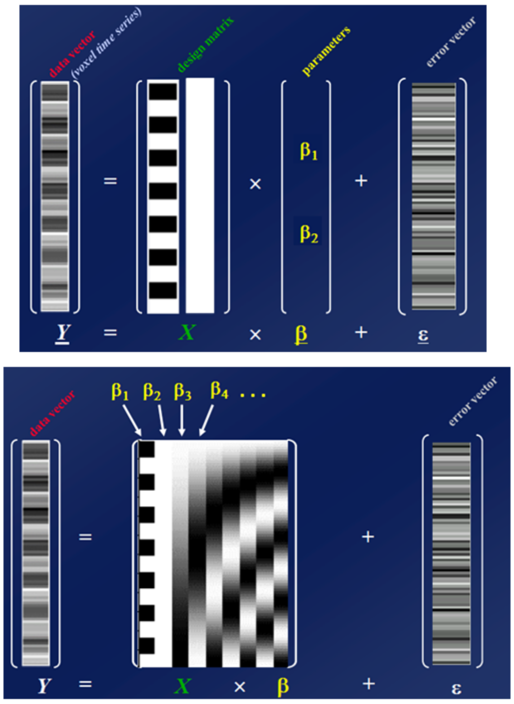

data vector
(voxel time series)
design matrix
parameters
error vector
=
×
+
β_1
β_2
$\underline{Y} = X \times \underline{\beta} + \underline{\varepsilon}$
data vector
β_1 β_2 β_3 β_4 ...
error vector
=
+
$Y = X \times \beta + \varepsilon$

Figure 4.4: Top, is shown the simplest GLM model as in Figure 4.2. Bottom, is shown a model with more than two regressors. In both cases the firsts regressors refers to the experiment design and the second to the base signal. In the bottom figure, the rest of regressors are the result of the convolution function (section 4.2.3). Usually, movement regressors (section 4.2.4) are in the last columns from the design matrix. [15]

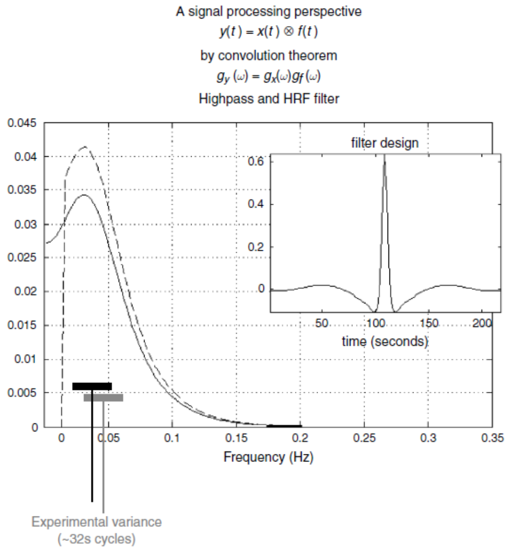


Figure 4.5: Transfer function of a canonical Haemodynamic Response Function (HRF), with (broken line) and without (solid line) the application of a highpass filter. This transfer function corresponds to the spectral density of a white-noise process after convolution with the HRF and places constraints on the frequencies that survive haemodynamic convolution. The top right figure is the filter expressed in time, corresponding to the spectral density that obtains after convolution with the HRF and highpass filtering. [3]

realistic because there are more than two regressors; in the same Figure, bottom the regressors, except the first two, express these convolutions.

By the convolution theorem, the frequency structure of experimental variance should therefore be designed to match the transfer function of the HRF. The corresponding frequency profile of the transfer function is shown in Figure 4.5 (solid line). It can be seen that frequencies around 0.03 Hz are optimal, corresponding to periodic designs with 32-second periods (i.e. 16-second epochs).

The BOLD impulse response in this voxel loads mainly on the canonical HRF, but also significantly on the temporal and dispersion derivatives. The canonical HRF combined with time and dispersion derivatives comprise a basis set, as the shape of the canonical response to conform the haemodynamic response that is commonly observed. The incorporation of the derivative terms allow for variations in subject-to-subject and voxel-to-voxel responses. The time derivative allows the peak response to vary by plus or minus a second and the dispersion derivative allows the width of the response to vary by a similar amount.

4.2.4 Movement Regressors and Others

It is possible to add regressors to the design matrix without going through the convolution process described above. An important example is the modelling of movement-related effects, which are reflected in the processing. Movement expresses itself in the data directly, and not through any haemodynamic convolution, these are added directly as explanatory variables in the usual way.
Because of these movement regressors and other regressors needed to adjust the signal, the design matrix is presented with many regressors (columns), but only the columns at the beginning are effects of interest to explain the behaviour of the activation the voxel during specific task.

4.2.5 Fitted and Adjusted Data

Once the model is estimated two new concepts appear: these are *fitted* and *adjusted data*. *Adjusted* data is the measure response of rCBF (section 4.1) obtained by the GLM, and *fitted data* is the same without the error term. Both, fitted and adjusted data can be expressed using just one (or more) regressors model or a linear combination of them (using contrast), this gives the response obtained by the model just for a specific condition of the experiment.

In the right side of Figure 4.3 are showed two types of signals which are estimated from the model. The *fitted data*, dashed line, is obtained from the model only with the design matrix and parameters. Otherwise, *adjusted data*, solid line, is the same as before, but also adding the estimated error ϵ.

4.3 SPM data analysis

4.3.1 Uni-voxel Model

As is said in section 4.2, GLM gives a model for the response variable rCBF measure for a voxel. Here is described the matrix formulation of the GLM. Rewriting 4.2, for one voxel (*uni-voxel* or univariate regression) we have this set of simultaneous equations which correspond to T different time points:

$$\begin{aligned}
&Y_1 = x_{11}\beta_1 + \ldots + x_{11}\beta_l + \ldots + x_{1L}\beta + \epsilon_1 \\
&\vdots = \vdots \\
&Y_t = x_{t1}\beta_1 + \ldots + x_{tl}\beta_l + \ldots + x_{tL}\beta + \epsilon_t \\
&\vdots = \vdots \\
&Y_T = x_{T1}\beta_1 + \ldots + x_{Tl}\beta_l + \ldots + x_{TL}\beta + \epsilon_T
\end{aligned}$$

This has an equivalent matrix form:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_t \\ \vdots \\ Y_T \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1l} & \dots & x_{1L} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{t1} & \dots & x_{tl} & \dots & x_{tL} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{T1} & \dots & x_{Tl} & \dots & x_{TL} \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_l \\ \vdots \\ \beta_L \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_t \\ \vdots \\ \epsilon_T \end{pmatrix}$$

At that point all is applied to only one voxel, but it is easy to see the same model can be applied to a set of voxels (for example, the set of voxels corresponding to a ROI). We call this set of voxels *mega-voxel.* In this case, each component Y_t from the response variable Y, corresponds to the average of all values from the mega-voxel at time t.

4.3.2 Multi-voxel Model

As is stated in [39], the *multi-voxel* (or multivariate) fMRI regression model, is a generalization of the univariate regression model from a single voxel to N voxels. The model $Y = XB + E$, represents a model where Y is an $n \times p$ matrix of the measure response where nth column, Y^n, the response for the nth voxel; X is the $T \times L$ design matrix containing regressors; B is the matrix of regressor coefficients with nth column, β^n, being for the nth voxel; and E is the matrix of error terms where nth column, E_n, are the errors for the nth, alternatively, the rows of E, $\epsilon_j \sim N(0, \Sigma)$ where Σ is the spatial covariance matrix between voxels.

Let us suppose in the fMRI the whole brain is formed by N voxels. Following, the matrix formulation for N voxels is presented in T time points.

$$\begin{pmatrix} Y_1^1 & \dots & Y_1^N \\ \vdots & \ddots & \vdots \\ Y_t^1 & \dots & Y_t^N \\ \vdots & \ddots & \vdots \\ Y_T^1 & \dots & Y_T^N \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1l} & \dots & x_{1L} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{t1} & \dots & x_{tl} & \dots & x_{tL} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{T1} & \dots & x_{Tl} & \dots & x_{TL} \end{pmatrix} \cdot \begin{pmatrix} \beta_1^1 & \dots & \beta_1^N \\ \vdots & \ddots & \vdots \\ \beta_l^1 & \dots & \beta_l^N \\ \vdots & \ddots & \vdots \\ \beta_L^1 & \dots & \beta_L^N \end{pmatrix} + \begin{pmatrix} \epsilon_1^1 & \dots & \epsilon_1^N \\ \vdots & \ddots & \vdots \\ \epsilon_t^1 & \dots & \epsilon_t^N \\ \vdots & \ddots & \vdots \\ \epsilon_T^1 1 & \dots & \epsilon_T^N \end{pmatrix}$$

Notice that the design matrix, and therefore the number of parameters β, is the same as in previous section and the model corresponds to a single subject which was exposed to specific task (Figure 4.6). This is because the design matrix encodes and quantifies our knowledge about how the expected signal was produced.

In some cases, when is needed to analyse the response

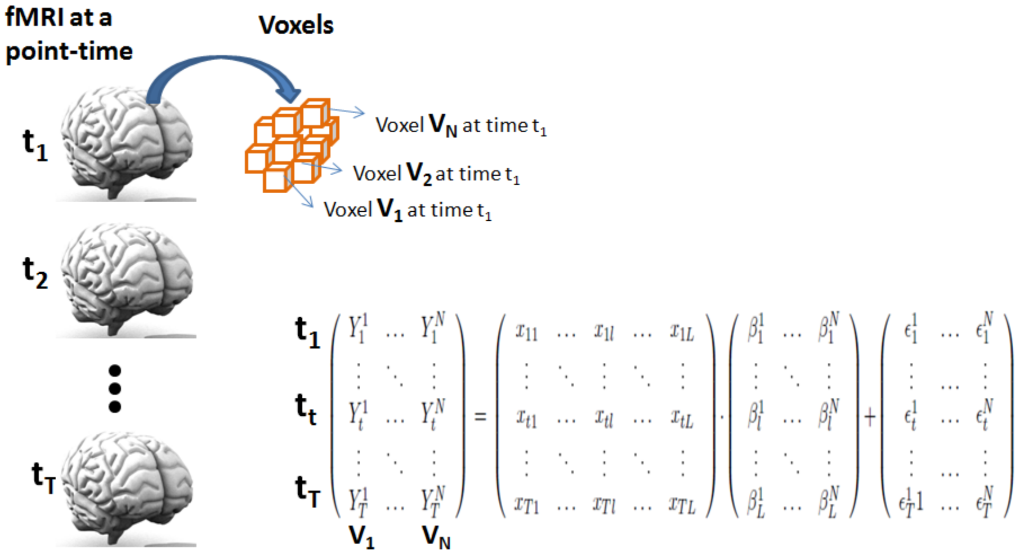

fMRI at a point-time
Voxels
t_1
t_2
t_T
Voxel $\mathbf{V}_N$ at time t_1
Voxel $\mathbf{V}_2$ at time t_1
Voxel $\mathbf{V}_1$ at time t_1
$$\begin{array}{cc} & \\ t_1 & \\ & \\ t_t & \\ & \\ t_T & \\ & \end{array}\begin{array}{c}\begin{pmatrix} Y_1^1 & \dots & Y_1^N \\ \vdots & \ddots & \vdots \\ Y_t^1 & \dots & Y_t^N \\ \vdots & \ddots & \vdots \\ Y_T^1 & \dots & Y_T^N \end{pmatrix} \\ \mathbf{V}_1 \qquad \mathbf{V}_N \end{array} = \begin{pmatrix} x_{11} & \dots & x_{1l} & \dots & x_{1L} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{t1} & \dots & x_{tl} & \dots & x_{tL} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{T1} & \dots & x_{Tl} & \dots & x_{TL} \end{pmatrix} \cdot \begin{pmatrix} \beta_1^1 & \dots & \beta_1^N \\ \vdots & \ddots & \vdots \\ \beta_l^1 & \dots & \beta_l^N \\ \vdots & \ddots & \vdots \\ \beta_L^1 & \dots & \beta_L^N \end{pmatrix} + \begin{pmatrix} \epsilon_1^1 & \dots & \epsilon_1^N \\ \vdots & \dots & \vdots \\ \epsilon_t^1 & \dots & \epsilon_t^N \\ \vdots & \dots & \vdots \\ \epsilon_T^1 1 & \dots & \epsilon_T^N \end{pmatrix}$$

Figure 4.6: This is an example of multivariate approach where the model is estimated for all the voxels from whole brain (*multivariate model*)

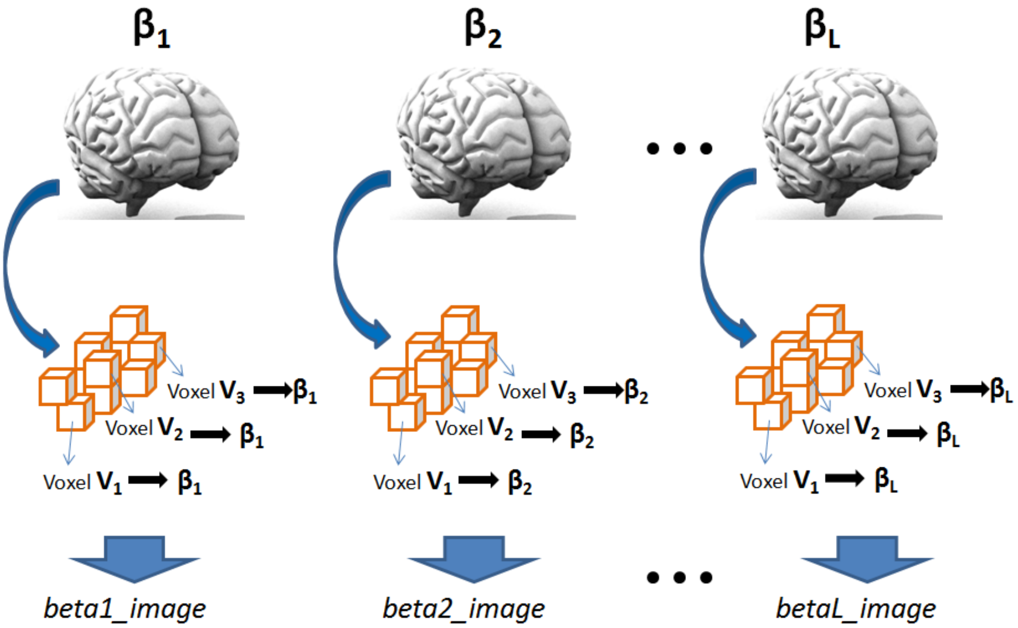
4.3.3 First Level Analysis

Given the *specification* of the GLM design matrix and after the *estimation* of the GLM parameters, the objective of the *first level analysis* is to analyse the presence of the defined activation pattern in a single voxel or mega-voxel.

In other words, SPM performs a statistical test to see how the signal voxel (noise removed) follow a activation pattern which is defined by the main regressors in the design matrix and the contrasts.

The model information is stored in *beta* and *contrast images*. Since this is a multivariate model we have only one design matrix and the same number of parameters β for each voxel. A *beta image* has exactly the same number of voxels as one fMRI, and there is as many beta images as the number of β parameters from the model (L). Figure 4.7 shows a diagram to visualise what beta image is.

Each parameter β in the model is associated with a condition effect of interest (section 4.2.1). On the other hand, a contrast (section 4.2.2) allows to compare the difference in responses between some conditions. A *contrast image* contains the contrast values, and there is as many contrast images as the number of defined contrasts (C).



β_1
β_2
β_L
Voxel $V_3 \rightarrow \beta_1$
Voxel $V_2 \rightarrow \beta_1$
Voxel $V_1 \rightarrow \beta_1$
Voxel $V_3 \rightarrow \beta_2$
Voxel $V_2 \rightarrow \beta_2$
Voxel $V_1 \rightarrow \beta_2$
Voxel $V_3 \rightarrow \beta_L$
Voxel $V_2 \rightarrow \beta_L$
Voxel $V_1 \rightarrow \beta_L$
beta1_image
beta2_image
betaL_image

Figure 4.7: This is a visual explanation for the definition of *beta images*

4.3.4 Second Level analysis

The objective of the *second level analysis* is to perform a subject group comparison. It uses the statistics summary from the first-level approach where contrast images from each subject are used as measures of subject responses. These are entered into the second level analysis as the new dependent variables and are analysed across subjects (*group analysis*).

Let us suppose we want to distinguish between two groups.Actually, we want to see if there is a significant difference between them. This is a special case of the GLM. The two-sample t-test assumes $Y_{qj} \overset{idd}{\sim} \mathrm{N}{\sim} \, (\mu_q, \sigma^2)$, for $q = 1, 2$, and assesses the same null hypothesis $H_0 : \mu_1 = \mu_2$, where μ_1 and μ_2 are the statistic measures from two different groups. The index j indexes the observation point in both groups.

The standard statistical way of writing the model is:

$$Y_{qj} = x_{qj1}\mu_1 + x_{qj2} + \epsilon_{qj}$$

The q from μ_q indicates the group. Here the regressors indicate the group membership, where x_{qj1} indicates whether observation Yqj is from the first group, in which case we have the value 1 when $q = 1$, and 0 when $q = 2$. Similarly, $x_{qj2} = \begin{cases} 0 & \text{if } q = 1 \\ l & \text{if } q = 2 \end{cases}$

Here a two sample t-test is used because there are two groups, and the null hypothesis is $H_0{=}\mu_1 = \mu_2$. But in case we want to see if there are differences between many groups, then the null hypothesis is all the groups have the same, which is $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_Q$, where Q is the number of groups. This tests can be solved using a specific F-statistic.

The way to evaluate the t-statistic (or F-statistic) is not explained in detail because second level analysis is not need to be used.

4.4 fMRI in ADHD Problem

4.4.1 Univariate Approach

In order to understand our contribution, first let us review the steps presented in [1]:

There is activation in IFG and VStr ROI?:
They perform a first-level analysis in order to assess the validity of the paradigms to activate the selected ROIs. Using SPM they performed one-sample t-test of IFG and VStr activity separately. They find, beta values (for “nogo>go” and “win>control” contrasts) extracted from individual ROIs are significantly different from zero in the right IFG as well as in the VStr during Go-NoGo and MID task respectively. This means that there is activation (specific activation pattern) in the target regions.

Whole brain group analysis:
After the first-level analysis, the main contrasts for each subject are entered into a second-level two sample t-test. With a p-value p<0.0001 (which is significant) the result revealed reduced activation in the patient group during reward anticipation in VStr region. Otherwise they do not detect any reduction of activity in IFG region in ADHD samples.

Standard ROI analysis:
They use the ROI as a mask to extract the beta values from the contrasts derived from the second-level analysis and it shows reduced VStr activity in the ADHD group analysis during the MID task. But they do not observe significant differences with regard to IFG activity.

4.4.2 Our Proposal: Multivariate Approach

Out aim is to investigate how can we use the data from first-level analysis performed in [1] and apply machine learning methods to perform the group analysis.

We suggest that through incorporation of machine learning tools into functional neuroimaging studies we will be able to identify unique patterns of variability in brain activation. Given fMRI data from specific tasks we want to deduce the behavioural brain activation and use it to classify between controls and patients.

In this project, we consider a classification problem: separating different groups of human subjects based on the observed fMRI time sequences. We contribute a comprehensive framework of spatially and temporally exploring fMRI data, and apply it to a challenging case of study: separating ADHD patients subjects from healthy controls based on their observed fMRI time sequences.

Basically it can be done by extracting some features from the model estimated and use them to build a classifier. We call it *multivariate classifier* because the variable we use to classify are: the beta parameters from the GLM; the linear combinations from the most important betas; and temporal information performed by GLM from specific contrast (fitted data).

The features are obtained as follows: First, a multi-voxel model for each subject will be created and therefore the design matrix and the beta parameters will be obtained, also the contrasts will be defined. This data, particularly beta parameters and contrasts, only from the voxels from the ROIs (IFS and VStr), will be used as a feature for a particular subject. Secondly, a uni-voxel model for just a selected ROI (mega-voxel) is generated and we compute the fitted data for the main contrasts, these are the temporal features for a subject.

Main characteristics of this approach (new paradigms):

- All features are obtained from the GLM estimated (without HRF-free) as in [5, 44, 37]

- We perform dimensionality reduction, feature selection and feature extraction, as is said in literature about machine learning approaches for fMRI data analysis.

- Basically the classifier tested is linear and non-linear SVM, because it is also used in the most of literature.

- The classifier must be able to perform a group analysis, it must be accurate in classify between patients and controls.

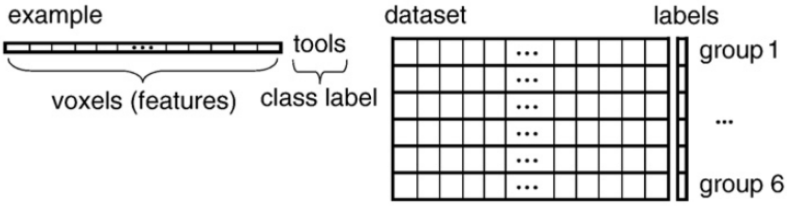


Figure 4.8: An example where features are voxels arrayed as a row vector (left) and a dataset is matrix of such row vectors (right). The column vector next to the dataset contain the label of each example (row).[46]

Actually, is expected to ratify machine learning approaches perform a good classifications between groups using the data from VStr during MID task. And also if we want to obtain a good classifier, using IFG for Go-NoGo task, which is able to distinguish between controls and patients.

4.5 Pattern Recognition System

4.5.1 Classification

As is stated in [46] a *classifier* is a function that takes the values of various features (independent variables in regression) in an example (the set of independent variable values) and predicts the class that example belongs to (the dependent variable). Each new example we want to classify or each example already classified is called *feature vector* and is represented by x. A *feature vector* with d features is represented by $x = (x_1, \dots, x_d)$. Given an example x its class label is denoted as y. All the examples are organized in a *dataset matrix*, where rows are examples and columns are features. In Figure 4.8 there is an illustration of a dataset matrix and a column vector which are the labels.

A classifier has a number of parameters that have to be learned from *training data*. A *training data* is a set of examples reserved for this purpose. The learned classifier is essentially a model of the relationship between the features and the class label in the training set. More formally, given a feature vector x, the classifier is a function f that predicts the label $y = f(x)$. Once trained, the classifier can be used to determine whether the features used contain information about the class of the example. This relationship is tested by using the learned classifier on a different set of examples, the *test data* (Figure 4.9). Intuitively, the idea is that, if the classifier truly captured the relationship between features and classes, it has to be able to predict the

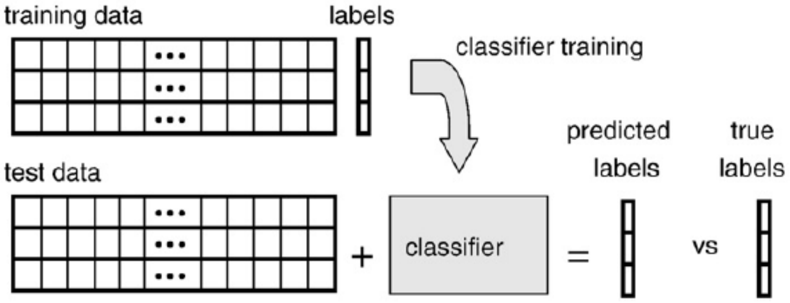


Figure 4.9: classifier is learned from the training set, examples whose labels it can see, and used to predict labels for a test set, examples whose labels it cannot see. The predicted labels are then compared to the true labels and the accuracy of the classifier, the fraction of examples where the prediction was correct, can be computed.[46]

classes of examples that hasn't seen before.

We will denote the training and test sets by X_{train} and X_{test}, matrices with respectively n_{train} and n_{test} examples as their rows, and their respective labels by the column vectors y_{train} and y_{test}.

4.5.2 Dimensionality Reduction

In most of the cases, if the subjects are patients, we don't dispose of a big sample because of the accessibility. So, maybe because we extract a many features for one subject or may be because we have a very few number of subjects, generally there are many more features than examples. As a consequence, the classifier will be a function that can classify well the examples in the training set, but not the examples in the test set. This phenomenon is called *overfitting*. That is why it may be advantageous to reduce the number of features considered. The crucial issue to keep in mind is that the choice of features at this stage must not depend on the labels that will be used to train a classifier. There are two alternatives to reduce the number of features which a classifier has to consider:

- *Feature Selection*:
  In this case the most relevant k features are selected ($k < d$), and the remaining $d - k$ features are ignored.

$$x = (x_1, \dots, x_d) \rightarrow x = (x_1, \dots, x_k), \text{ where } d < k$$

- *Feature Extraction*:
  An alternative path to reduce the number of features is the dimensionality reduction. This is applied to the entire dataset matrix; they transform the original feature space into a new low dimensional feature space. $x = (x_1, \dots, x_d) \rightarrow z = (z_1, \dots, z_d)$, where $d << k$ This yields a new dataset matrix with the same number of rows but a reduced number of columns. However, it is not at all guaranteed to improve results, partially because most dimensionality reduction techniques ignore class labels in their criteria.

Principal Component Analysis

The feature extraction method we use here in this work is Principal Component Analysis (PCA). The goal of PCA is to reduce the dimensionality of the data while retaining as much as possible of the variation present in data set.
The way is to find a new representation (basis) able to filter the noise and reveal hidden dynamics. Using the eigenvectors from the covariance matrix, a new coordinate system is formed and it's defined by the significant axis. Extract the eigenvectors with a high eigenvalue ad it leads to compress data: reduce the dimension of feature vector in a new representation and keeps the as much a possible the variability.

4.5.3 Support Vector Machine

In machine learning, *Support Vector Machines* (SVM) are supervised learning models[6] with associated learning algorithms that analyse data and recognize patterns, used for classification. Given a set of training examples labelled, the basic SVM train an algorithm and build a model that assigns new examples into one category or the other. It takes a set of input data and predicts, for each given input, which of two possible classes forms the input, making a non-probabilistic binary linear classifier.

A generally idea about how SVM works is that, given a d-dmensional feature space, SVM try to separate the points (features vectors in that space) with a $(d-1)$-dimensional hyperplane. There are many hyperplanes that might classify the data; one reasonable choice as the best hyperplane is the one that represents the largest separation, or margin[7], between the two classes. So SVM choose the hyperplane whose distance from it to the nearest

[6] Models where the examples are already labelled, i.e. the class of any example is known

[7] Margin is the perpendicular distance from the hyperplane to the closes samples

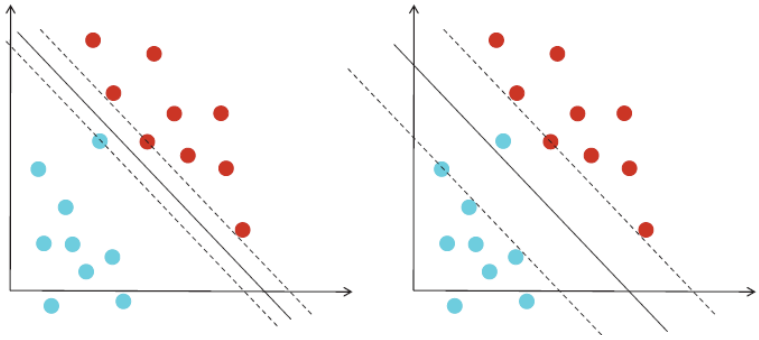


Figure 4.10: The left figure shows the typical SVM Linear classification, and the right figure, is exactly the same but in this case a misclassified is allowed, so the margin is softer.

data point on each side is maximized. In Figure 4.10 (left side) the continuous line is the hyperplane which maximise the distance between the two classes, and the distance is the same between one class to the hyperplane and between to the hyperplane to the other class.

Even if a decision boundary exactly separate the data, if the data has noise and outliers, a soft margin decision boundary that ignores a few data points is better, as is shown in Figure 4.10. The parameter C is called *cost of constrain violation* and it permit the algorithm to misclassify some of the data points without affecting the final result.

Since it is a linear classifier, which find the best hyperplane, it is called *Linear support vector machine* (LinearSVM). But sometimes it is easy to compute kernels which correspond to complex large-dimensional feature spaces, for example the radial basis kernel, in this case the classifier is called *Radial Basis Function Support Vector Machine* (RbfSVM). This kernel has the following formula:

$$k(x, x') = exp(-\frac{\|x - x'\|^2}{\gamma^2}), \gamma > 0 \tag{4.10}$$

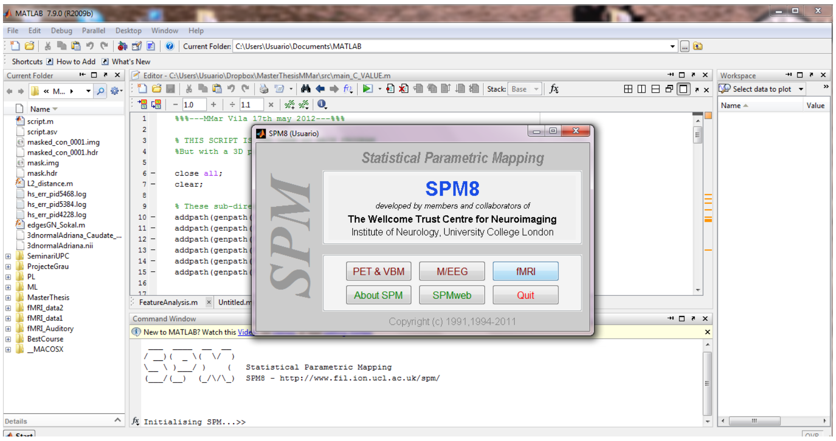
Thus, changing the γ parameter we obtain different classifiers

The choice of this method is due to the good performance in real world applications, and also the most common classifier used in the literature. Also it is advantageous to use SVM because of its Computational efficiency and because our problem is not very high dimensional space.

Chapter 5

Technical Development

5.1 Development Platforms



```
%%%---MMar Vila 17th may 2012---%%%

% THIS SCRIPT IS
%But with a 3D p

close all;
clear;

% These sub-dire
addpath(genpath(
addpath(genpath(
addpath(genpath(
addpath(genpath(
addpath(genpath(
addpath(genpath(
```

```
 ___  ____  __  __
/ __)(  _ \(  \/  )
\__ \ )___/ )      (   Statistical Parametric Mapping
(___/(__)  (_/\/\_)   SPM8 - http://www.fil.ion.ucl.ac.uk/spm/

Initialising SPM...>>
```

Figure 5.1: SPM 8 interface

On one hand Matlab Version 7.9 (R2009b) was used to develop machine learning algorithms. Matlab allows matrix manipulations, plotting of functions and data and implementation of algorithms. It is is very simple to extract individual rows, columns and submatrices using a very powerful indexing system. That makes Matlab a programming environment optimized for internal routines for matrix transformations very suitable for data analysis, visualization, and algorithm development.

To construct the Support Vector Machine (SVM) classifier was used OSU SVM Vesrion 3.00 which is a SVM toolbox for the MATLAB environment. The toolbox can be used to create models for regression and classification using SVM[1].

On the other hand, to do the analysis of fMRI data we used the Matlab toolbox SPM 8. Figure 5.1 hows the SPM 8 interface available. The SPM software package is a suite of Matlab and it has been designed for the analysis of brain imaging data sequences. Specifically this version is designed for the analysis of fMRI, PET, SPECT, EEG and MEG. SPM is made freely available to the neuro-imaging community, to promote collaboration and a common analysis scheme across laboratories. The software represents the implementation of the theoretical concepts of Statistical Parametric Mapping (section 4.1)) in a complete analysis package.

MarsBaR (MARSeille Boîte À Région d'Intérêt) is a region of interest toolbox for SPM which provides routines for region of interest analysis. Features include region of interest definition, combination of regions of interest with simple algebra, extraction of data for regions with and without SPM preprocessing (scaling, filtering), and statistical analyses of ROI data using the SPM statistics machinery.

All programs were executed in a machine equipped with an I5 processor and 4GB of RAM memory.

5.2 Preprocessing

All preprocessing steps are done with SPM [15].

- *Slice Timing*:
  Correct differences in image acquisition time between slices. The slice order argument that specifies slice acquisition order is a vector of S numbers, where S is the number of slices per volume. Each number refers to the position of a slice within the image file. The order of numbers within the vector is the temporal order in which those slices were acquired. The function provided by SPM use this vector to correct differences in slice acquisition times. The correction is necessary to make the data on each slice correspond to the same point in time. This routine "shifts" a signal in time to provide an output vector that represents the same (continuous) signal sampled starting either later or earlier. This is accomplished by a simple shift of the phase of the sines that make up the signal. Recall that a Fourier transform allows for a representation of any signal as the linear combination of sinusoids of different frequencies and phases. A constant to the phase of every frequency, shifting the data in time. Then a filter (called Shifter) is

[1] *http://sourceforge.net/projects/svm/*

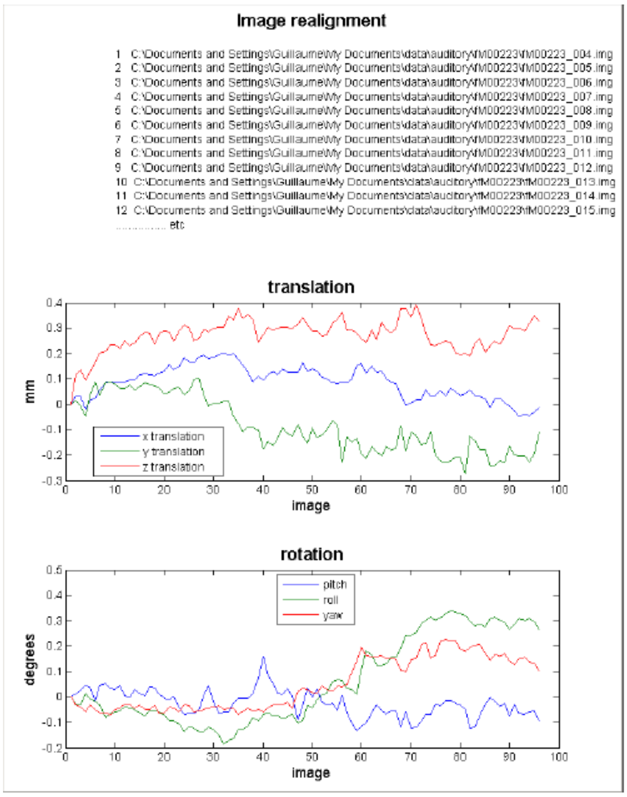


Image realignment

```
1   C:\Documents and Settings\Guillaume\My Documents\data\auditory\fM00223\fM00223_004.img
2   C:\Documents and Settings\Guillaume\My Documents\data\auditory\fM00223\fM00223_005.img
3   C:\Documents and Settings\Guillaume\My Documents\data\auditory\fM00223\fM00223_006.img
4   C:\Documents and Settings\Guillaume\My Documents\data\auditory\fM00223\fM00223_007.img
5   C:\Documents and Settings\Guillaume\My Documents\data\auditory\fM00223\fM00223_008.img
6   C:\Documents and Settings\Guillaume\My Documents\data\auditory\fM00223\fM00223_009.img
7   C:\Documents and Settings\Guillaume\My Documents\data\auditory\fM00223\fM00223_010.img
8   C:\Documents and Settings\Guillaume\My Documents\data\auditory\fM00223\fM00223_011.img
9   C:\Documents and Settings\Guillaume\My Documents\data\auditory\fM00223\fM00223_012.img
10  C:\Documents and Settings\Guillaume\My Documents\data\auditory\fM00223\fM00223_013.img
11  C:\Documents and Settings\Guillaume\My Documents\data\auditory\fM00223\fM00223_014.img
12  C:\Documents and Settings\Guillaume\My Documents\data\auditory\fM00223\fM00223_015.img
............... etc
```

Figure 5.2: Example of realignment of fMRI data. [15]

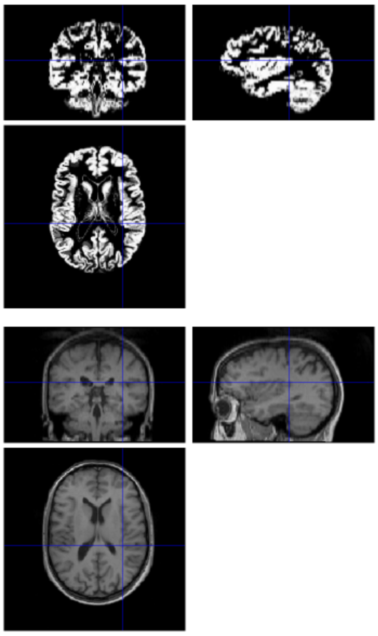


Figure 5.3: Example of segmenation of fMRI data. [15]

used and the signal will be convolved to introduce the phase shift. It is constructed explicitly in the Fourier domain. In the time domain, it may be described as an impulse (delta function) that has been shifted in time the amount described by a *time shift*. The correction works by lagging (shifting forward) the time-series data on each slice using *sinc-interpolation*.

- *Realignment*:
  This routine realigns a time-series of images acquired from the same subject using a least squares approach and a 6 parameter spatial transformation (rigid transformation). The first image in the list specified by the user is used as a reference to which all subsequent scans are realigned. The reference scan does not have to the first chronologically and it may be wise to chose a "representative scan" in this role. The aim is primarily to remove movement artefact in fMRI time-series (or more generally longitudinal studies). Given all fMRI data SPM write realigned images based with respect to *mean image*. Then SPM plot the estimated time series of rigid transformations, translations and rotations, shown in Figure 5.2. A rigid-body transformation (in 3D) can be parametrized by three translations and three rotations about the different axes (in total 6 parameters as we said above). These variables

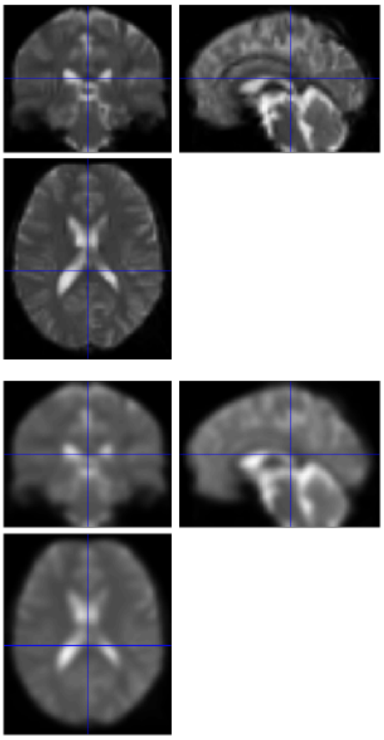


Figure 5.4: Functional image (top) and 6mm-smoothed functional image (bottom). [15]

can be used as regressors when fitting GLMs. This allows movements effects to be discounted when looking for brain activation.

- *Coregistration*:
  Image registration is the process of transforming different sets of data into one coordinate system. Data may be data from different times. Registration is necessary in order to be able to compare or integrate the data obtained from these different measurements. An image similarity measure quantifies the degree of similarity between intensity patterns in two images. The choice of an image similarity measure depends on the modality of the images to be registered. Common examples of image similarity measures include mutual information [2]. SPM implement a coregistration between the structural and functional data that maximises the mutual information.

- *Segmentation*:
  SPM will segment the structural image using the default tissue probability maps as priors. Note that this module needs the images to be roughly aligned with the tissue probability maps before you begin. This model also includes parameters that account for image intensity non-uniformity. Optimal results can be obtained reducing the number of Gaussians per class or increasing the sampling distance. This model also includes parameters that account for image intensity non-uniformity. Then, SPM will create gray and white matter images and bias-field corrected structural image. Figure 5.3 shows the gray matter image along with an original structural.

- *Normalise*:
  Thismodule spatially normalises images into a standard space defined by some ideal model or template images. The template images supplied with SPM and approximate to the space described in the atlas of Talairach and Tournoux (1988). The transformation can also be applied to any other image that has been coregistered with these scans. Generally, the algorithms work by minimising the sum of squares difference between the image which is to be normalised, and a linear combination of one or more template images. For the least squares registration to produce an unbiased estimate of the spatial transformation, the image

[2] This is a probability term. Given two random variables is a quantity that measures the mutual dependence of the two random variables. In our case, given a reference image (for example, a brain scan), and a second image which needs to be put into the same coordinate system as the reference image, this image is deformed until the mutual information between it and the reference image is maximized.

contrast in the templates (or linear combination of templates) should be similar to that of the image from which the spatial normalisation is derived. The registration simply searches for an optimum solution. If the starting estimates are not good, then the optimum it finds may not find the global optimum.

- *Smooth*:
  This is for smoothing (or convolving) image volumes with a Gaussian kernel of a specified width. It is used as a preprocessing step to suppress noise and effects due to residual differences in functional and gyral anatomy during inter-subject averaging. In Figure 5.4 there is an smoothed functional image.

5.3 Feature Description

We want to use machine learning algorithms to classify between two groups, so we need to define the feature vectors to describe the data.

We consider fMRI data from two groups, ADHD patients and controls and the GLM is obtained as well as the first level analysis. Our objective is to get feature from the GLM and model a classifier able to distinguish between both groups given these features as input. We use two kinds of feature as input. On one hand, features extracted explicitly from the GLM, such as *betas* and *contrast* (section 4.2.2). And, on the other hand, features extracted using the GLM for a specific contrast (the main contrast for the task), such fitted data (section 4.2.5), which contain temporal information of the data.

Thus, we consider three different feature vectors, which are *Betas*, *Contrasts* and *Fitted Data*, so we train three different classifiers. For the three type of features, the feature values are computed by the average on the specific ROI. Summarizing, a subject can be characterized for three types of feature vectors, as follows:

- Feature vector *Betas*: $\bar{x}_\beta = (x_1, x_2, \ldots, x_L)$, where L is the number of betas and $x_l, l \in \{1, \ldots, L\}$ is the mean of all l-th beta for all the voxels from the ROI.

- Feature vector *Contrasts*: $\bar{x}_c = (x_1, x_2, \ldots, x_C)$, where C is the number of contrasts and $x_c, c \in \{1, \ldots, C\}$ is the mean of al c-th contrast for all the voxels from the ROI.

- Feature vector *Fitted Data*: $\bar{x}_f = (x_1, x_2, \ldots, x_T)$, where T is the number of time-points of the task and $x_t, t \in \{1, \ldots, T\}$ is the value at time

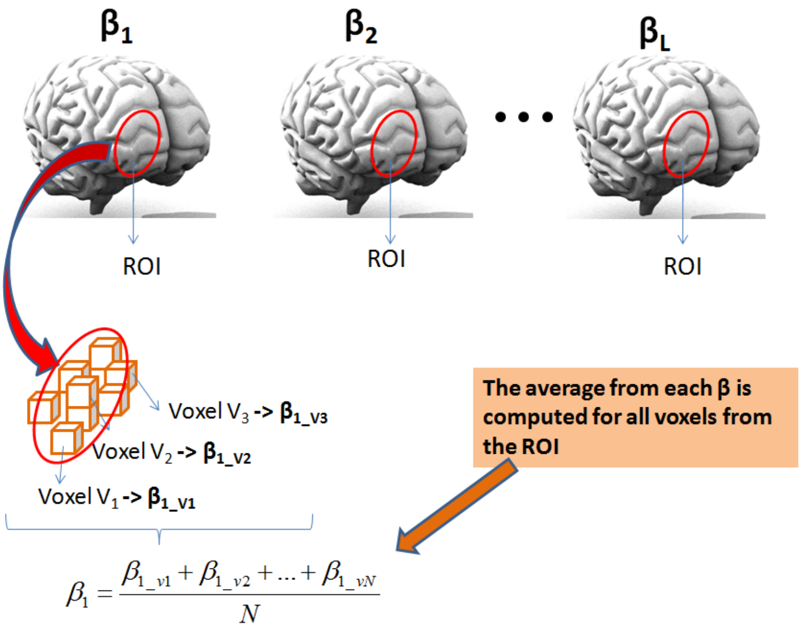

β_1
β_2
β_L
ROI
ROI
ROI
Voxel V_3 -> β_{1_v3}
Voxel V_2 -> β_{1_v2}
Voxel V_1 -> β_{1_v1}
The average from each β is computed for all voxels from the ROI
$\beta_1 = \frac{\beta_{1\_v1} + \beta_{1\_v2} + \ldots + \beta_{1\_vN}}{N}$

Figure 5.5: This is a visual description for the beta experiment.

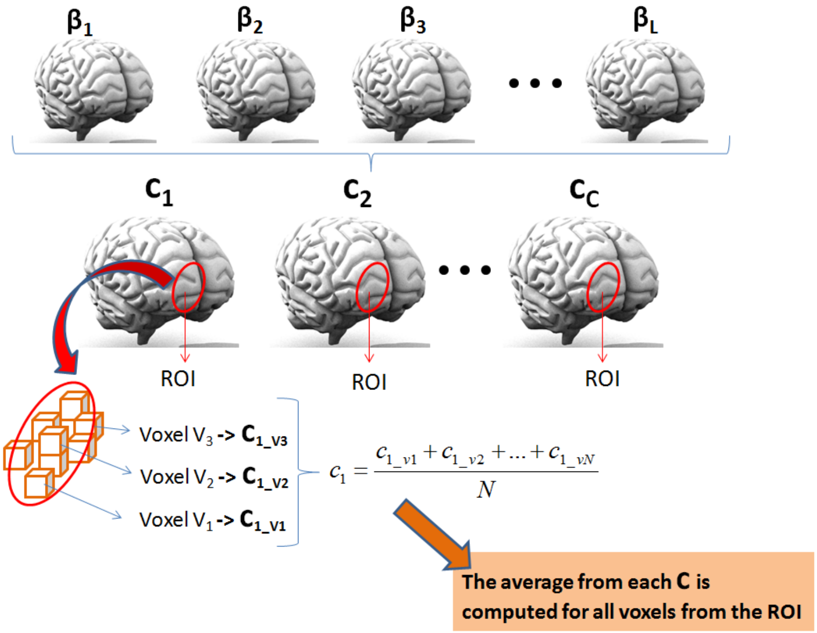

β_1
β_2
β_3
β_L
C_1
C_2
C_C
ROI
ROI
ROI
Voxel V_3 -> C_{1_v3}
Voxel V_2 -> C_{1_v2}
Voxel V_1 -> C_{1_v1}
$c_1 = \frac{c_{1\_v1} + c_{1\_v2} + \ldots + c_{1\_vN}}{N}$
The average from each C is computed for all voxels from the ROI

Figure 5.6: This is a visual description for the contrasts experiment.

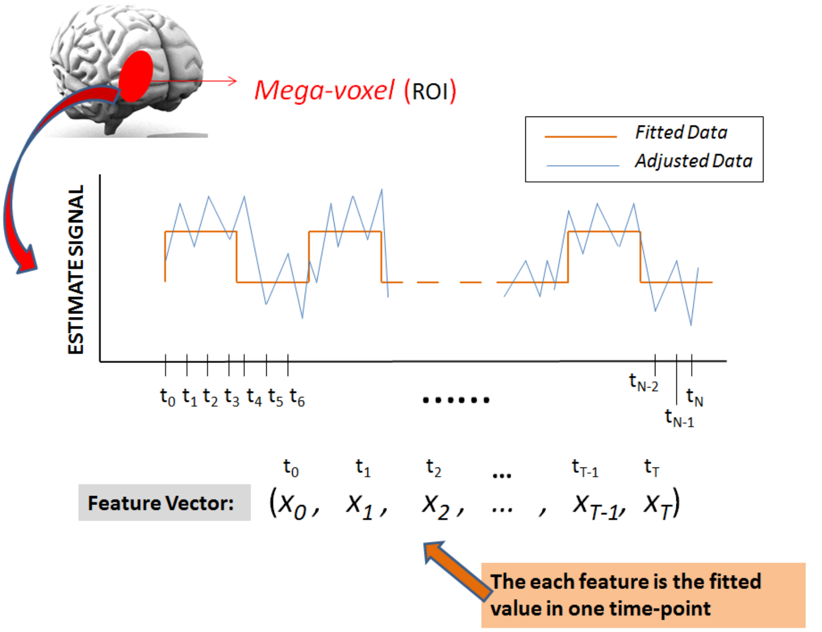


Figure 5.7: This is a visual description for the fitted data experiment.

t of the mega-voxel (corresponding to the ROI) obtained as the measure response, i.e. the column vector Y from the GLM model (Equation (4.1)), for a particular contrast.

In Figures 5.5, 5.6 and 5.7 there is a visual representation of the three types of features.

5.4 Efficacy of the Classifier

The most commonly used measure of how well a classifier does on the test set is its *accuracy*. This is simply the fraction of feature vectors in the test set for which the correct label was predicted. This is

$$Accuracy = \frac{\#\text{examples predicted correctly}}{n_{test}} \tag{5.1}$$

In case it is a binary classification test there are other statistical measures or rates, these are *sensitivity* and *specificity*. In a binary classification one class is called the "positive" and the other, the "negative". Then *sensitivity* measures the proportion of actual positives which are correctly identified and *specificity* measures the proportion of negatives which are correctly identified.

These statistical measures are evaluated with the following measures:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{5.2}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5.4}$$

Where TP= “True Positive” correspond to the number of examples belonging to the positive class and are predicted correctly; TN= “False Negative” correspond to the number of examples belonging to the negative class and are predicted correctly; FP= “False Positive” correspond to the number of examples belonging to the positive class and are predicted as a negative and

FN= “True Negative” correspond to the number of examples belonging to the negative class and are predicted as a positive.

5.5 Validation

In the illustrative example of Figure 4.9 the data set is divided into halves, one it’s used for training and the other for testing. In order to obtain the most accurate classifier we would like to train a classifier as much as possible. But we cannot train and test in the same data if we want to obtain a useful estimate of the classifier’s accuracy. The procedure proposed is called *cross-validation.* This process consists in divide all the data set in equal parts. One part is used for the test set and the others are used as a training set. The process is repeated for each part and then the averages of the rates called in section 5.4 are computed in order to evaluate the efficacy of the method An important consideration using these methods are that training data in each fold must contain examples of all classes, otherwise the classifier for that fold will not be able to predict the absent ones.

This method has to variants *Leave One Out* (LOO) and *k-cross-validation*:

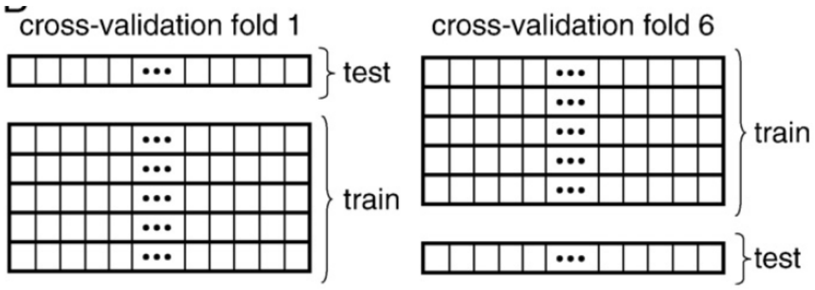


Figure 5.8: During cross-validation each of 6 groups of examples takes a turn as the test set while the rest serve as the training set.

- *LOO*:
  This method leaves one example out, train with the rest of the examples and a prediction is made of this example. This is repeated for each example and then the accuracy and the other rates are computed averaging all the predictions made for all the examples. In practice, leaving each example out can be computationally expensive because the number of classifiers trained is the same as the number of examples, an alternative is proposed bellow.

- *k-cross-validation*:
  Where k is the number of parts into which dataset is divided, and the process is exactly the same. Common choices are $k = 10$ or $k = 5$, corresponding to leaving out 10% or 20% of the examples on each fold. In the illustrative example shown in Figure 5.8 the data set is divided in six groups and each one is used in turn as the test set in cross validation, with the remaining groups used as the training set.

Chapter 6

Experiments

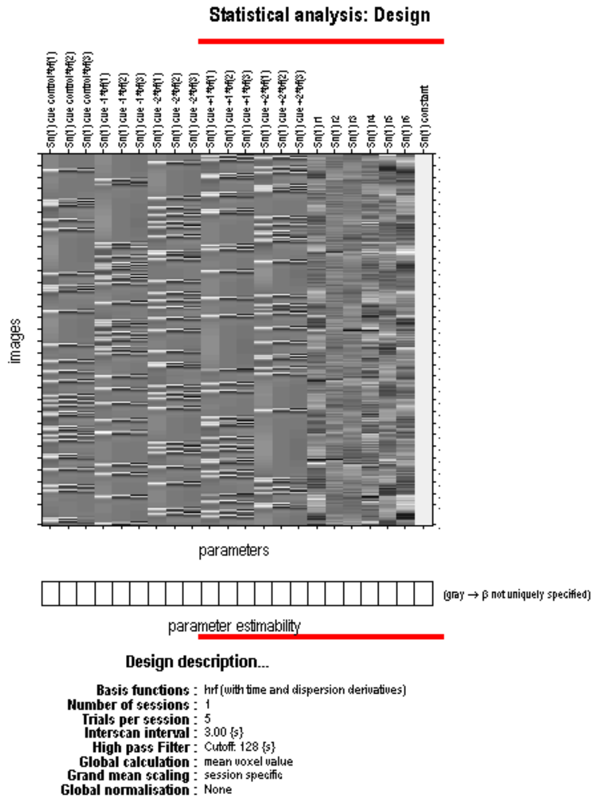
6.1 Data Acquisition

As is stated in [1] during the fMRI acquisition, the subjects performed two tasks: a *Go-NoGo* task (6 minutes and 59 seconds) and a *Monetary Incentive Delay* (*MID*) task (14 minutes and 31 seconds). See Section 6.1.3 for details of MID and Go-NoGo tasks. The order of presentation was counterbalanced across participants to avoid undesirable confounding effects.

The series of temporal interpolation were performed over each voxels' time course using sinc functions. (slice timing from section 5.2). Subsequently, spatial interpolation was applied to correct for head motion, using parameters derived from a rigid body transformation. Specialists checked the individual translation and rotation movement parameters, which did not exceed a value of 4 mm/4 degrees for any of the subjects. The EPI[1] images were subsequently smoothed by imposing an 8-mm FWHM isotropic kernel on the space domain. The normalization step was omitted from the preprocessing of the fMRI data to avoid a potential bias in the results arising from ADHD-associated volumetric and shape alterations in the IFG and the VStr. Therefore, they created ROI of these areas in each individual's anatomical space.

In order to delimitate the IFG, the following steps were applied to the data. For each subject, the T1-weighted image was co-registered to the mean functional EPI, and subsequently segmented into grey matter, white matter, and cerebrospinal fluid partitions. Then, the inverse parameters created

[1] Echo Plannar Iaging (EPI) is a popular technique for rapid acquisition and is used to extract our fMRI data. The main reason to use this technique is, for all studies, the shape of an individual's fMRI data will be more similar to their anatomical scan and this improves the quality of the normalization leading to improved group level statistics throughout the brain.



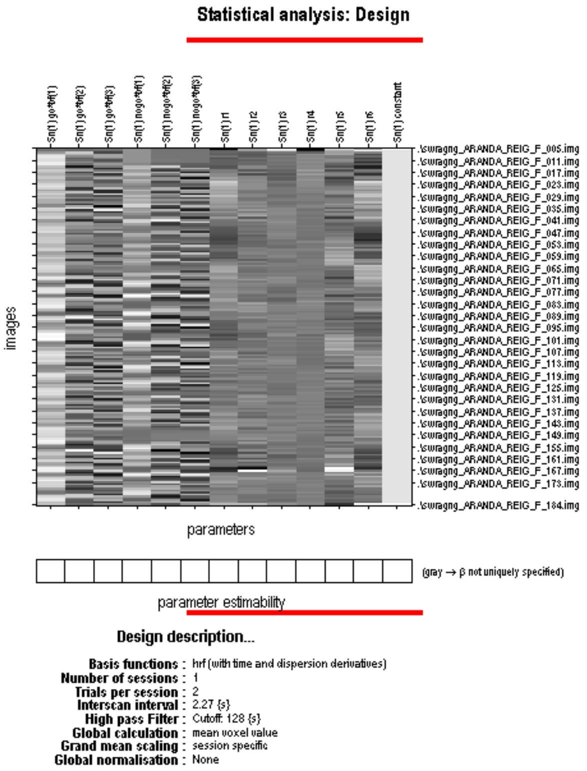
Design description...

Basis functions : hrf (with time and dispersion derivatives)
Number of sessions : 1
Trials per session : 5
Interscan interval : 3.00 {s}
High pass Filter : Cutoff: 128 {s}
Global calculation : mean voxel value
Grand mean scaling : session specific
Global normalisation : None

Figure 6.1: This is the design matrix for the MID task. As is said in section 5.2 about spatial preprocessing, specifically realignment, some regressors refer to movement effects, these are the penultimate 6 in this design matrix

in the segmentation step were applied to anatomical ROIs representing the left and right IFG. In addition, the individual grey matter partitions were imposed on the created ROIs to restrict the regions to grey matter tissue. These ROIs were then individually revised by a neuroradiologist and manually adjusted in MRIcroN. The VStr ROIs were manually delineated under the guidance of an expert neuroradiologist. The criteria for Vstr demarcation are described in detail in [47]. To ensure intra-rater reliability of the VStr-ROI, we repeated the segmentation of 10 VStr (5 left and 5 right) and calculated the overlap coefficient (intersection/union), resulting in an average ROI overlap of 0.71 (0.65 for the left VStr and 0.76 for the right VStr), and the Interclass Correlation Coefficient absolute agreement, which resulted in an index of 0.84 (0.89 for the left VStr and 0.80 for the right).

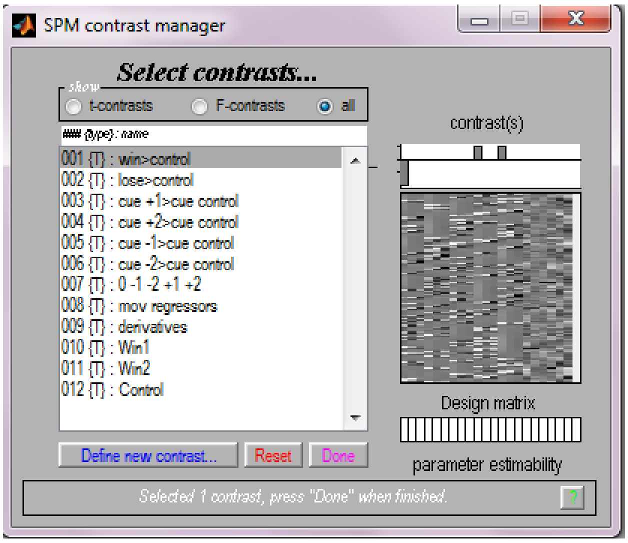
Subsequently, voxel-wise changes in BOLD response across the conditions


Statistical analysis: Design
Sn(1) go*bf(1)
Sn(1) go*bf(2)
Sn(1) go*bf(3)
Sn(1) nogo*bf(1)
Sn(1) nogo*bf(2)
Sn(1) nogo*bf(3)
Sn(1) r1
Sn(1) r2
Sn(1) r3
Sn(1) r4
Sn(1) r5
Sn(1) r6
Sn(1) constant
images
.\swragng_ARANDA_REIG_F_005.img
.\swragng_ARANDA_REIG_F_011.img
.\swragng_ARANDA_REIG_F_017.img
.\swragng_ARANDA_REIG_F_023.img
.\swragng_ARANDA_REIG_F_029.img
.\swragng_ARANDA_REIG_F_035.img
.\swragng_ARANDA_REIG_F_041.img
.\swragng_ARANDA_REIG_F_047.img
.\swragng_ARANDA_REIG_F_053.img
.\swragng_ARANDA_REIG_F_059.img
.\swragng_ARANDA_REIG_F_065.img
.\swragng_ARANDA_REIG_F_071.img
.\swragng_ARANDA_REIG_F_077.img
.\swragng_ARANDA_REIG_F_083.img
.\swragng_ARANDA_REIG_F_089.img
.\swragng_ARANDA_REIG_F_095.img
.\swragng_ARANDA_REIG_F_101.img
.\swragng_ARANDA_REIG_F_107.img
.\swragng_ARANDA_REIG_F_113.img
.\swragng_ARANDA_REIG_F_119.img
.\swragng_ARANDA_REIG_F_125.img
.\swragng_ARANDA_REIG_F_131.img
.\swragng_ARANDA_REIG_F_137.img
.\swragng_ARANDA_REIG_F_143.img
.\swragng_ARANDA_REIG_F_149.img
.\swragng_ARANDA_REIG_F_155.img
.\swragng_ARANDA_REIG_F_161.img
.\swragng_ARANDA_REIG_F_167.img
.\swragng_ARANDA_REIG_F_173.img
.\swragng_ARANDA_REIG_F_184.img
parameters
(gray → β not uniquely specified)
parameter estimability

Design description...

```
        Basis functions : hrf (with time and dispersion derivatives)
     Number of sessions : 1
     Trials per session : 2
      Interscan interval : 2.27 {s}
        High pass Filter : Cutoff: 128 {s}
     Global calculation : mean voxel value
     Grand mean scaling : session specific
   Global normalisation : None
```

Figure 6.2: This is the design matrix for the Go-NoGo task. As is said in section 5.2 about spatial preprocessing, specifically realignment, some regressors refer to movement effects, these are the penultimate 6 in this design matrix.



SPM contrast manager

Select contrasts...

show: ○ t-contrasts ○ F-contrasts ● all

{type} : name

001 {T} : win>control
002 {T} : lose>control
003 {T} : cue +1>cue control
004 {T} : cue +2>cue control
005 {T} : cue -1>cue control
006 {T} : cue -2>cue control
007 {T} : 0 -1 -2 +1 +2
008 {T} : mov regressors
009 {T} : derivatives
010 {T} : Win1
011 {T} : Win2
012 {T} : Control

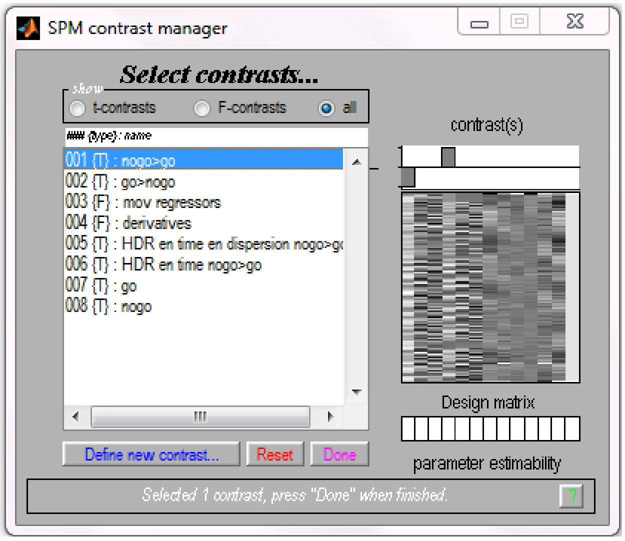
Selected 1 contrast, press "Done" when finished.

Figure 6.3: *Contrast Manager*. It displays the design matrix in the right panel and lists specified contrast in the left panel. Most of the contrasts are have been specified automatically. Here in this case are showed all the contrast for the MID task, and the main contrast "win>control" is selected.

were assessed for each subject, according to the general linear model. The Go-NoGo paradigm was applied to extract an indication of the inhibition response, and time courses for the "nogo" and "go" trials were introduced into the model. For the MID task, which was employed to assess reward anticipation, the onset times of "win cue" and "control cue" comprised the regressors of interest.

The regressors of interest were convolved with the canonical haemodynamic response function implemented in SPM, and optimal parameter estimates were computed using a least-squares function. Time and dispersion derivatives were also included in the model, as were the individual translation and rotation parameters in order to account for residual effects of movement. In Figure 6.1 and 6.2 is shown the design matrix for each task obtained from the model.

Finally, the linear contrasts "nogo>go" and "win>control" were applied to estimate effect sizes, as is shown in *Contrast Manager* from Figure 6.3 and 6.4 respectively. For each subject, the first-level models were then imported into Marsbar, and the individual ROIs representing the IFG and VStr in the subject's anatomical space were applied to these results. For each of the ROIs, we extracted the beta values for the contrasts "nogo>go" and "win>control".



{type} : name

001 {T} : nogo>go
002 {T} : go>nogo
003 {F} : mov regressors
004 {F} : derivatives
005 {T} : HDR en time en dispersion nogo>g
006 {T} : HDR en time nogo>go
007 {T} : go
008 {T} : nogo

Selected 1 contrast, press "Done" when finished.

Figure 6.4: *Contrast Manager*. It displays the design matrix in the right panel and lists specified contrast in the left panel. Most of the contrasts are have been specified automatically. Here in this case are showed all the contrast for the Go-NoGo task, and the main contrast "nogo>go" is selected.

6.1.1 Participants

Forty-six right-handed adult males (23 with ADHD and 23 healthy controls) were included in the study. Eight subjects (4 ADHD and 4 controls) were omitted from the analysis, due to problems understanding the tasks,to other complications occurring in either of the fMRI paradigms, or for neurological reasons. All the subjects were evaluated by a team of psychologists and psychiatrists from Vall d'Hebron Hospital. All ADHD subjects fulfilled the diagnostic criteria for ADHD and had never received any pharmacological treatment for their condition. ADHD diagnosis was based on the Diagnostic and Statistical Manual of Mental Diseases, Fourth Edition, Test Revised (DSM-IV TR).

6.1.2 fMRI Acquisition Parameters

The MRI images were obtained in a GE 1.5T scanner, equipped with a standard quadrature radiofrequency coil. A vacuum pillow was placed inside the coil in order to restrict the subject's head movement. For anatomical reference, a T1-weighted pulse sequence was employed with the following parameters: TR 11.5, TE 4.2, matrix 256 × 256 × 96, FA 15, slice thickness 1.6. Functional volumes were acquired using a T2*-weighted gradient echo

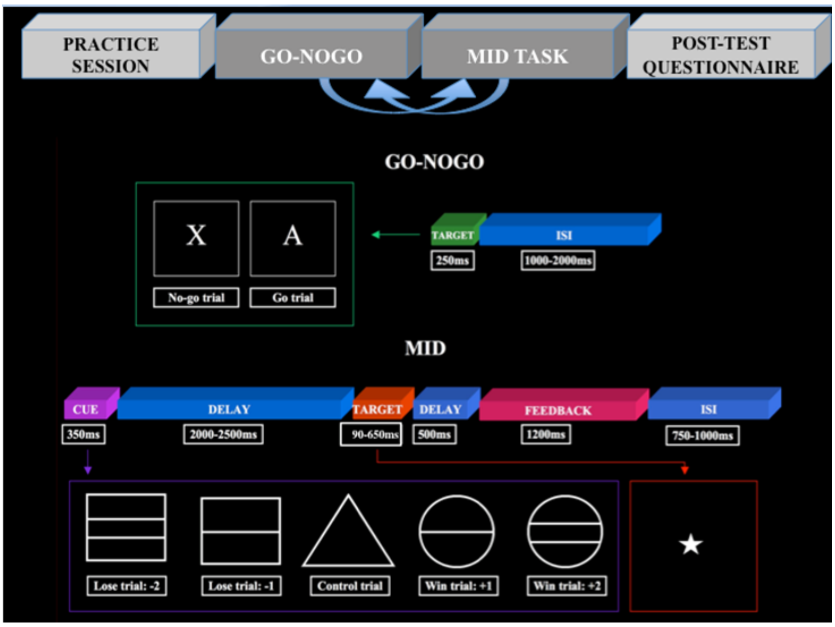


Figure 6.5: This figure depicts the timeline of the experimental procedure as well as the duration and different stimuli presented in each of the tasks. Light grey boxes represent measures obtained outside the scanner and dark gray boxes indicate the two fMRI paradigms measured inside the scanner [1].

sequence. For the MID task, the acquisition parameters were: TR = 3,000 ms; TE = 60 ms, FA = 90°, FOV = 30 cm, GAP = 0.5 and a matrix size of 64 × 64 × 30. The Go-NoGo paradigm was acquired using the following parameters: TR = 2,275 ms, TE = 60 ms, FA = 90°, FOV = 30 cm, GAP = 0.5 and a matrix size of 64 × 64 × 23.

6.1.3 fMRI Procedure

During the fMRI acquisition, the subjects participated in two event-related fMRI paradigms. The order of presentation was counterbalanced across subjects.

The Go-NoGo task was similar to the ones used in previous studies comprising the presentation of individual letters on a screen. The subjects were instructed to press the button when a letter appeared (go trials), but withhold their response when an “X” was presented (nogo trials). The stimulus duration was 250 ms, and each stimulus was followed by a random interstimulus interval between 1,000 and 2,000 ms. The total number of trials, on average, was 225. The percentage of go trials was set to 70% [48, 49]. A higher percentage of go trials relative to no-go trials has been demonstrated to enhance the MRI signal for the nogo trials, by increasing the preponderance of the press response and consequently the difficulty of inhibiting it [48].

To assess reward anticipation, we used a version of the MID task similar to those employed in previous studies [11, 12]. See Figure 6.5 for an illustration of the timeline of the experimental procedure for MID task. For this task, trials involved the presentation of a cue for 350 ms, followed by a variable delay of between 2,000 and 2,500 ms, and then a target with a duration of between 90 and 650 ms. The subjects were required to press the button before the target disappeared from the screen; target duration was adjusted to produce success in around 66% of the cases (short target duration between 90 and 120 ms; long target duration between 550 and 650 ms). The cues comprised symbolic signs indicating trials with the possibility to win money ($+1/+2$ euro), to lose money ($-1/-2$ euro), or to keep the same amount regardless of the performance (control trials). The total number of trials averaged 150, and each of the five conditions ($+1/+2/-1/-2$/control) was presented for 20% of the trials. The target was followed by a delay (500 ms) and then a feedback screen (1,200 ms), depicting the amount gained or lost by the subject in this trial and the total quantity earned so far. See Supporting Information for a detailed description of the task and the timeline of the experimental procedure.

6.2 Feature Analysis

Features defined in Section 5.3 are extracted for both tasks, MID and Go-NoGo, and using the fMRI data from 38 subjects: 19 controls and 19 patients. The design matrix in figures 6.1 and 6.2 show the betas and Contrast manager in figures 6.3 and 6.4.

- *MID task.* In this case the ROI is the Left and Right VStr
  - *Beta Experiment*

    There are 22 betas, so there are 22 features for each ROI (left and Right).
    Thus the feature vector will be $\bar{x}_\beta = (x_1, x_2, \ldots, x_{44})$.

    Also we want to do the same but only with main betas, this means discarding movement regressors and the constant.

There are 15 main betas, so there are 15 features for each ROI (Left and Right). Thus the feature vector will be $\bar{x}_\beta = (x_1, x_2, \ldots, x_{30})$.

  - *Contrasts Experiment*
    There are 12 contrasts, so there are 12 features for each ROI (left and Right).
    Thus the feature vector will be $\bar{x}_c = (x_1, x_2, \ldots, x_{24})$.
    Then, for some subjects all the values for one beta are NaN, we remove the subject from the data. Thus we have 17 controls and 16 patients.

  - *Contrasts Experiment*

    * There are 286 time points contrasts, so there are 286 features for each ROI (left and Right).
      Thus the feature vector will be $\bar{x}_f = (x_1, x_2, \ldots, x_{572})$.

    * We also believe it is useful to consider the fitted data from what SPM consider as a Effects Of Interest (EOI).
      It has the same characteristics as before.

- *Go-NoGo task* In this case the ROI is the Left and Right IFG.

  - *Beta Experiment*

    There are 13 betas, so there are 13 features for each ROI (left and Right).
    Thus the feature vector will be $\bar{x}_\beta = (x_1, x_2, \ldots, x_{26})$.

    Also we want to do the same but only with main betas, this means discarding movement regressors and the constant.
    There are 12 main betas, so there are 15 features for each ROI (Left and Right). Thus the feature vector will be $\bar{x}_\beta = (x_1, x_2, \ldots, x_{24})$.

  - *Contrasts Experiment*
    There are 8 contrasts, so there are 8 features for each ROI (left and Right).
    Thus the feature vector will be $\bar{x}_c = (x_1, x_2, \ldots, x_{16})$.
    One control's ROI are missing so 21 vs 21.
    Finally: 42 subjects, this means all data is represented in a 42×16

matrix.

- *Fitted Data Experiment*
  - There are 180 time points contrasts, so there are 180 features for each ROI (left and Right).
    Thus the feature vector will be $\bar{x}_f = (x_1, x_2, \ldots, x_{360})$.
    One control's ROI are missing so 21 vs 21. Finally: 42 subjects, this means all data is represented in a 42×360 matrix.

  - We also believe it's useful to consider the fitted data from what SPM consider as a effects of interest (EOI).
    It has the same characteristics as before.

In order to see if a classification make sense with the selected features, following we show some graphics where the mean of each feature for all the controls and patients subjects is evaluated separately. The means are computed with normalized data. Normalization of the data is performed by subtracting the mean of all subjects and dividing by the standard deviation. Figures 6.6, 6.7 and 6.8 show the mean of the controls (in blue) and the patients (in red). Two graphics are presented for each kind of feature separating the right and the left ROI in the right and left side of the figure, respectively. This way allow us to notice if any hemisphere of ROI give more information than the other.

In Figure 6.6 we can appreciate a slight separation between the two classes, more prominent in the MID task than the Go-NoGo and no difference are noted between the right and the left ROI. Also in Figure 6.7, we can appreciate a slight separation between the two classes, more prominent in the MID task than the Go-NoGo and no difference are noted between the right and the left ROI. Finally, in Figure 6.8 in both cases we can appreciate a separation between the two classes more prominent in the left ROI than in the right ROI.

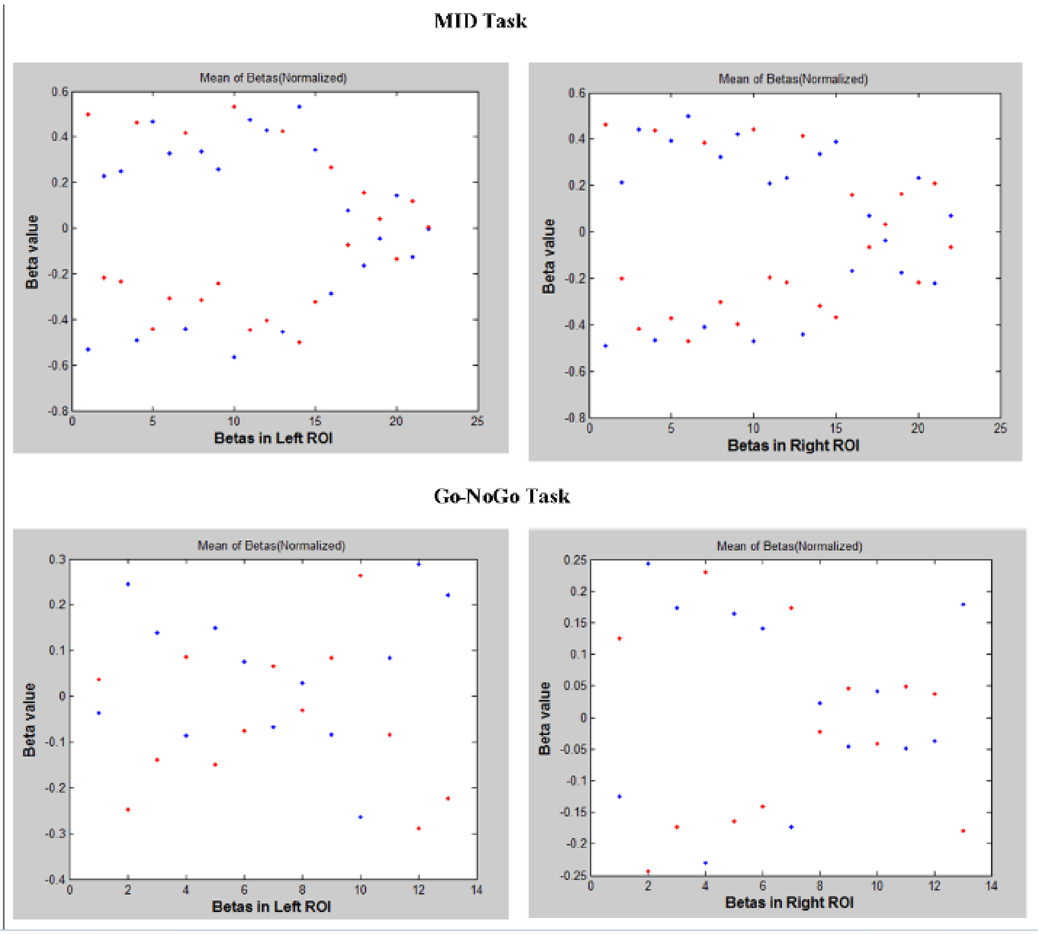


Figure 6.6: At the top, the mean data for betas features in MID task. At the bottom, the mean data for betas features in Go-NoGo task. The left side corresponds to the left ROI and the right side to de right ROI.

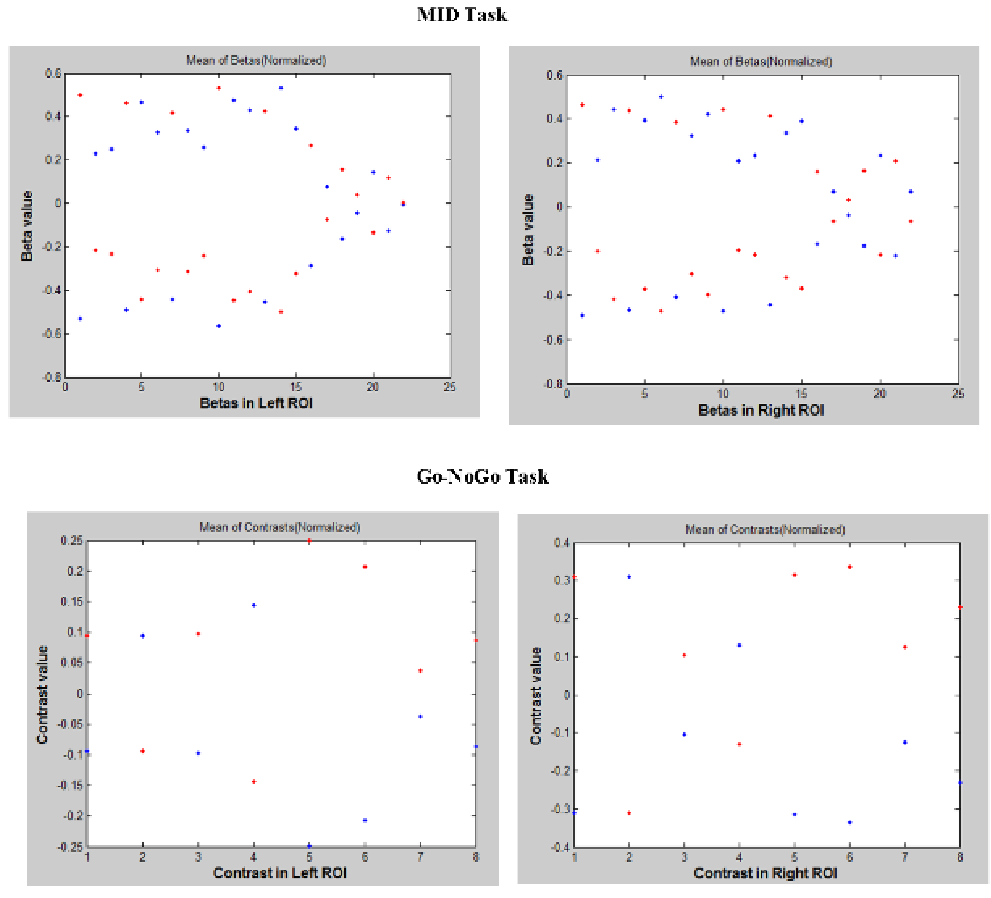


Figure 6.7: At the top, the mean data for contrasts features in MID task. At the bottom, the mean data for contrasts features in Go-NoGo task. The left side corresponds to the left ROI and the right side to de right ROI.

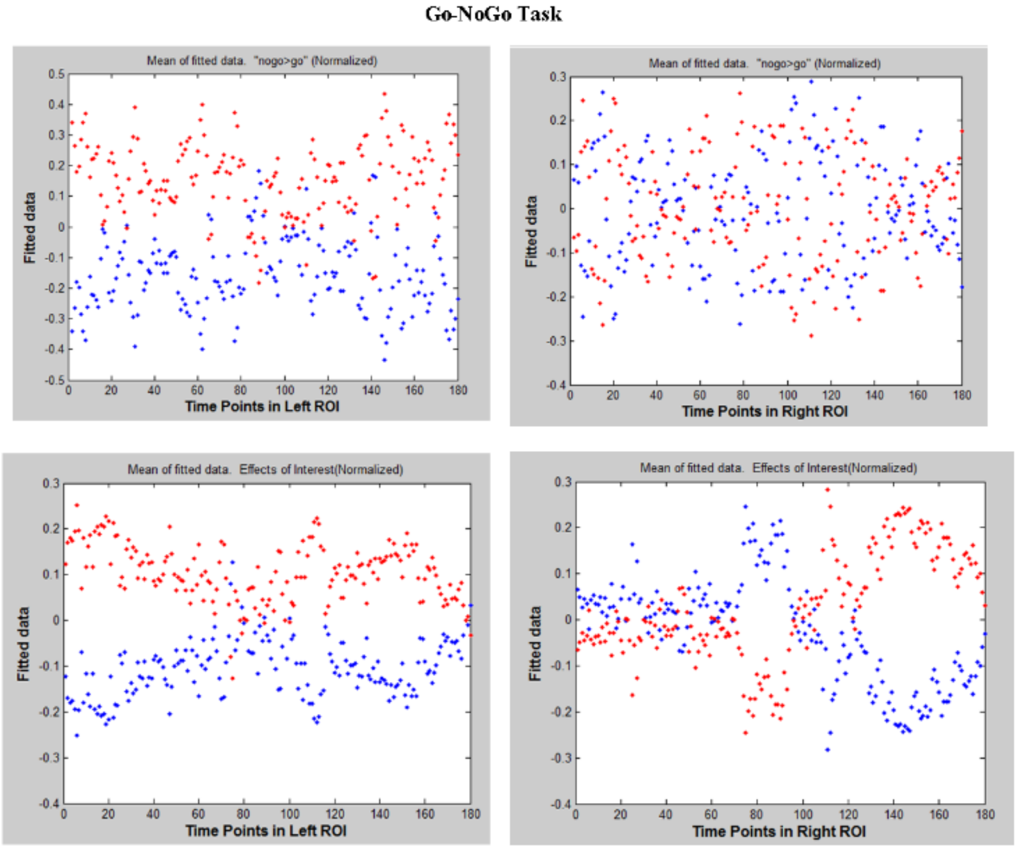


Figure 6.8: At the top, The mean data of fitted data features in Go-NoGo task. At the bottom, the mean data of effects of interest fitted data features in Go-NoGo taskin Go-NoGo task. The left side corresponds to the left ROI and the right side to de right ROI.

6.3 Parameters Setting

There is no previous work using this kind of features, so our contribution must focus on finding an appropriate method, adjusting the parameters, and selecting the correct features in order to find a good classifier able to find out if a subject is ADHD patients or not with certain reliability.

Our experiments are made to find the best model in order to classify between two classes, “patients” and “controls”. Since we classify between two classes it is a binary problem in which we use the variants of SVM classifier.

For the parameter setting of the classifier we use two methods: LOO, since we have just have not large amount of subjects, so it is not computationally expensive, and on the other hand 5-cross-validation to have bigger subsets.

In order to obtain an accurate validation of the classifier all the experiments are done with stratified data. Each rate (accuracy, specificity and sensitivity) is evaluated for each subset (subsets obtained in each iteration in LOO and 5-cross-validation) and then the average is computed in all parts, we use the value obtained to evaluate the efficacy of the method.

We have 4 different kinds of features: Betas, Contrasts, Fitted Data and EOI Fitted Data. In all of them we applied a feature dimensional reduction. Moreover, in case of Betas we applied a feature selection, discarding the features that correspond to movement regressors and the base, we call the selected features *main betas*. PCA is applied in all the features, because we think it's necessary because of the small number of subjects to train. In Fitted data and EOI fitted data we apply the two alternative reducing feature dimension. First we extract the features corresponding to the features from the side ROI (right or left) where the classes are further apart, in this case the Left ROI in case of VStr. But selecting the half of features (which corresponds to the one side ROI) is still a high number of feature dimension (note that in that case the features are time points, exactly 180 time points per a side ROI in Go-NoGo task) so it can still be necessary to reduce the dimension of feature, thus PCA is also applied. The number of eigenvectors we choose to form the new basis are those to keep the 90% of variance. This is calculated using the Portion of Variance (PoV) formula:

$$\text{PoV} = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_k + \ldots + \lambda_d} \tag{6.1}$$

For SVM classifier we set the C value parameter in different intervals and we focused in the intervals with best accuracys. The intervals we tried were:

$$c - values = [0.01 : 0.01 : 1]; [0.1 : 0.1 : 3] \text{ and } [1 : 1 : 10]$$

Also we tried methods with C values bigger than 10, but the accuracy reached was always the same or lower than the method with the other C values. In the intervals we focused the steps were 0.001.

In case of RbfSVM the parameter γ is also tried in three intervals and we focused in the intervals with best accuracys. These were:

$$\gamma - values = [0.001 : 0.001 : 1.0]; [0.1 : 0.01 : 3.0] \text{ and } [1 : 0.1 : 10];$$

For γ's bigger than 10 the accuracy obtained was always worse than the other values. In the intervals we focused the steps were 0.0001.

6.4 Results

In this section, we summarize the obtained results for the two tasks.

MID Task

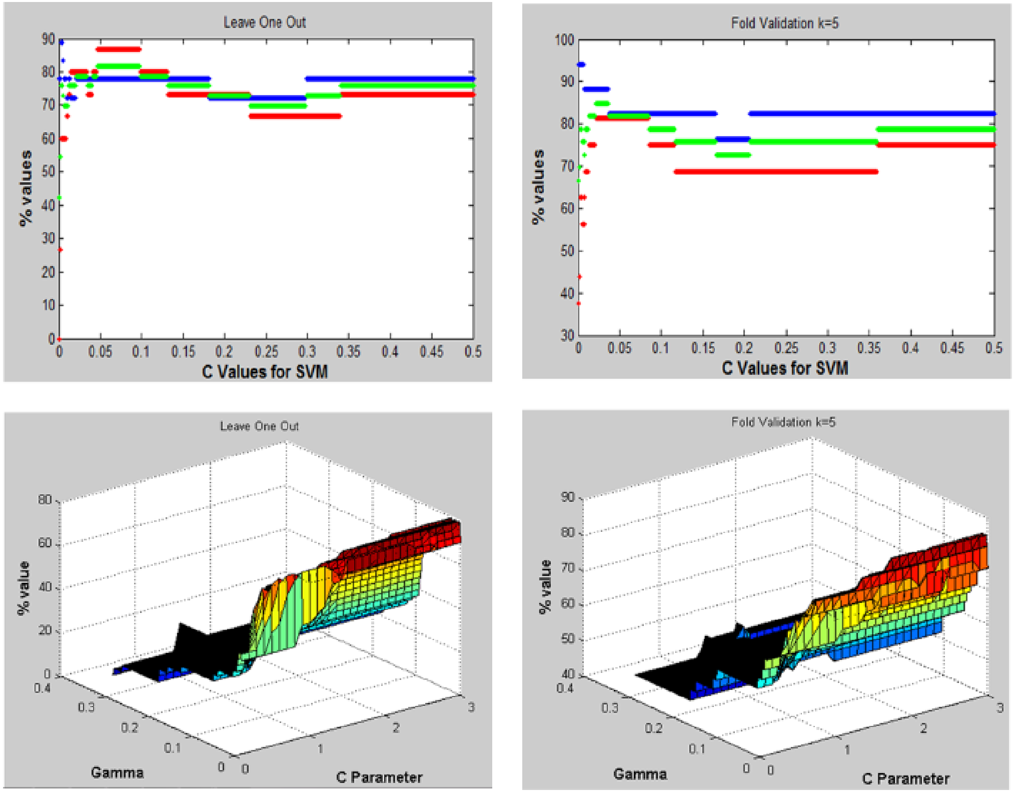


Figure 6.9: Best methods using betas features in MID task. Top: (sensitivity (red), Specificity (blue) and accuracy (green)) LinearSVM with C values from 0.001 to 0.5, at left the validation of the method with LOO and at right with 5-fold validation. Bottom: The accuracy, RbfSVM with C values from 1 to 10 and γ from 0.1 to 3, at left the validation of the method with LOO and at right with 5-fold validation

- *Betas*:

  Figure 6.9 shows the results for this experiment for different C values using LOO (left) and cross-validation (right); and using LinearSV (top) and RbfSM (bottom). The highest accuracy is around 85% of accuracy with both methods, LinearSVM and RbfSVM with γ less than 0.1. In case of the Linear SVM C value must be very small, around 0.5; and in case of RbfSVM it must be bigger than 2. PCA extract 3 features but the best method using these features just reach 55% of accuracy.

Selecting the main betas features the highest accuracy is slightly bellow 80% of accuracy with both methods, LinearSVM and RbfSVM with γ around 0.1. In the case of the Linear SVM C value must be between 0.1 and 0.2; and in case of RbfSVM it must be bigger than 0.5.

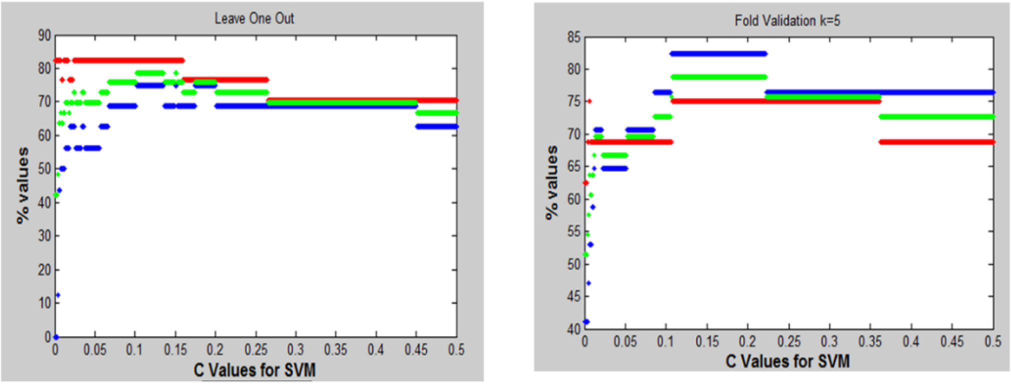


Figure 6.10: Best methods using contrast features in MID task (Sensitivity (red), Specificity (blue) and accuracy (green)). LinearSVM with C values from 0 to 0.5, at left the validation of the method with LOO and at right with 5-fold validation.

- *Contrasts*:

  Figure 6.10 shows the results for this experiment for different C values in LinearSVM using LOO (left) and cross-validation (right). The highest accuracy reach 80% of accuracy with LinearSVM and C value must be between 0.15. The highest accuracy with RbfSVM is around 70% with γ value around 0.1 and C value bigger than 8. PCA extract 2 features but the best method just reach 40% of accuracy.

- *Fitted Data*

  The highest accuracy reach with this data is around 60% of accuracy with LinearSVM using only data from the Left ROI and C value must be between 0.2 and 0.5. The other models performed have an accuracy around 50%.

- *Effects of Interest Fitted Data*

  These results are similar than before; the highest accuracy reach with this data is around 60% of accuracy with LinearSVM using only data from the Left ROI and C value must be between 0.1 and 0.5. The other models performed have an accuracy around 50%.

Go-NoGo Task

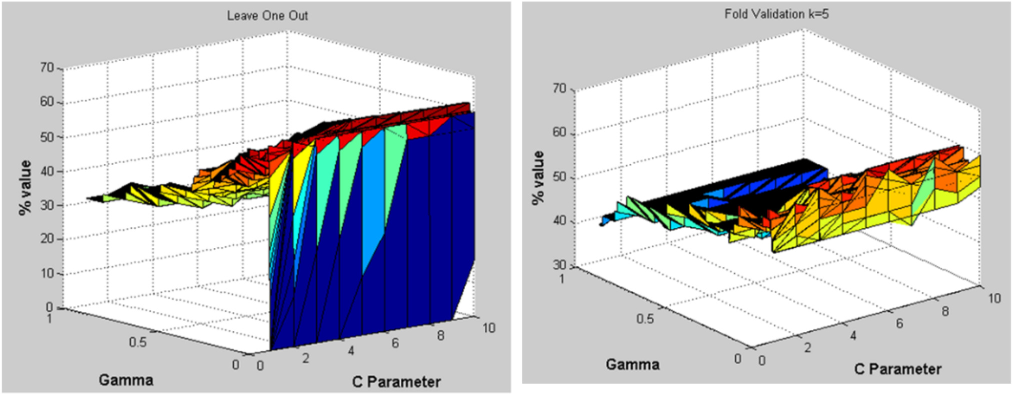


Figure 6.11: Accuracy of the best methods RbfSVM with C values from 1 to 10 and γ from 0.001 to 1, at left the validation of the method with LOO and at right with 5-fold validation.

- *Betas*:

  Figure 6.11 shows the results for this experiment for different C values in RbfSVM using LOO (left) and cross-validation (right). The highest accuracy reach 60% of accuracy with RbfSVM and C value bigger than 4 and a very smallγ, around 0.001. The highest accuracy with Linear SVM is around 55% . PCA extract 2 features but the best method just reach 40% of accuracy.

  Selecting the main betas the highest accuracy is slightly bellow 60% of accuracy with both methods, LinearSVM and RbfSVM with γ around 0.1. In case of the Linear SVM C value must be between 1.2 and 1.5; and in case of RbfSVM it must be bigger than 10.

- *Contrasts*:

  The highest accuracy is slightly bellow 60% of accuracy with LinearSVM method, using PCA and not. Using PCA, the best method is with a C value around 0.2 while without using PCA the best method is with a C values around 1. The RbfSVM methods don't reach the 50% of accuracy.

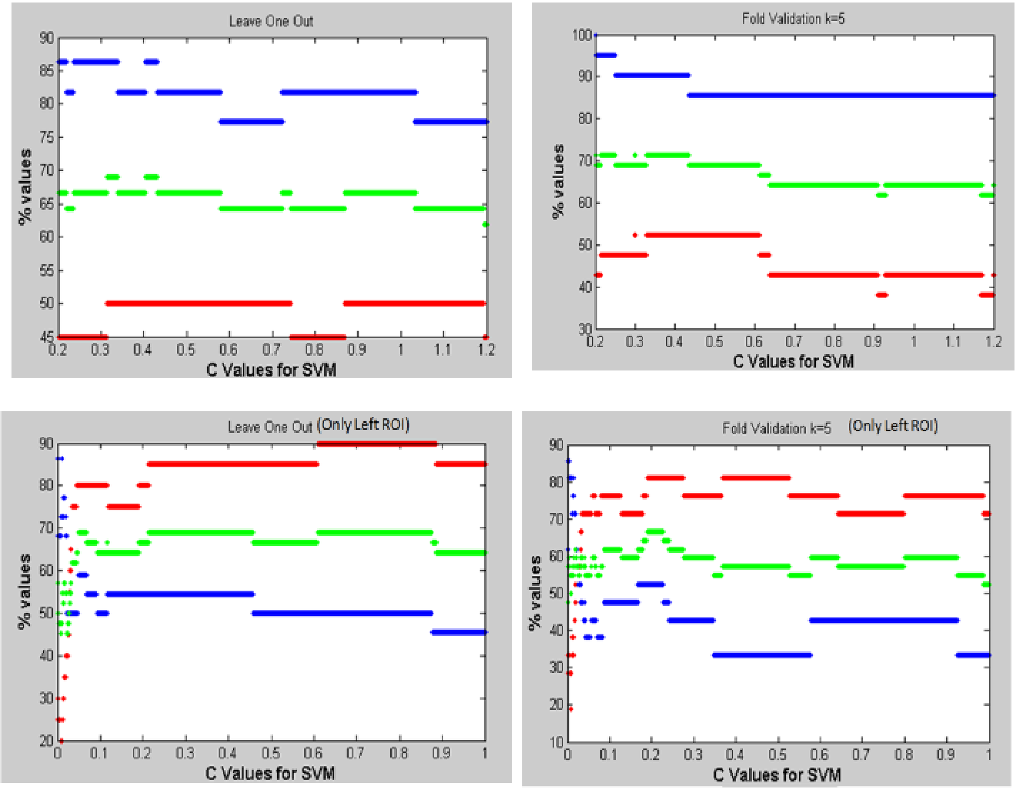


Figure 6.12: Rates of the best methods (Sensitivity (red), Specificity (blue) and Accuracy (green)). At the top LinearSVM using all features with PCA and C value from 0.2 to 1.2 (left LOO and right 5-fold-validation). At the bottom LinearSVM using the features only for the Left ROI with C value from 0 to 1 (left LOO and right 5-fold-validation).

- *Fitted Data*:

  Figure 6.12 shows the results for this experiment for different C values and dimensionality reduction; using LOO (left) and cross-validation (right); and using PCA (top) and feature selection (bottom). The highest accuracy is around 70% for two LinearSVM methods. One is using all features and PCA (obtaining 17 features) and with a C value around 0.4. The other one is using the features only for the Left ROI and the C value around 0.2.

  Another LinearSVM method reach an accuracy between 60% and 70% and it is using all the fatures and a C value around 0.03. Using the other methods the accuracy is around 60%.

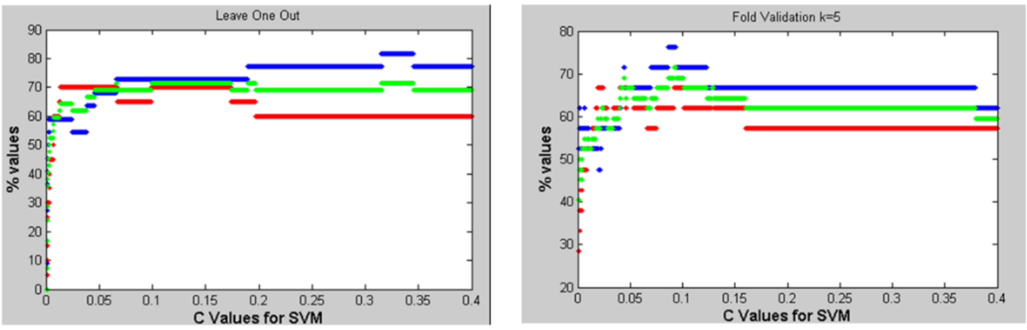


Figure 6.13: Rates of the best methods (sensitivity (red), Specificity (blue) and accuracy (green)). LinearSVM using the features only for the Left ROI with C value from 0 to 0.4 (left LOO and right 5-fold-validation).

- *Effects of Interest Fitted Data*:

  Figure 6.13 shows the results for this experiment for different C values in Linear SVM using LOO (left) and cross-validation (right). The higher accuracy is around 70% with the LinearSVM and C value around 0.1, using all features from the Left ROI. Also a high accuracy, between 60% and 70%, but never reach 70%, with LinearSVM using all features, from the whole ROI, with C value between 0.05 and 0.07. RbfSVM methods just reach the 60% of the accuracy using the features from the left ROI, for γ's from 0.2 to 0.4 and C values bigger than 1.5. In all methods using PCA the accuracy is around 50% or 55%.

6.5 Discussion

In general, we can see that the best classification results are obtained using betas as feature vector on the data from VStr ROI (MID task), with accuracies that reach 85%. We also find accurate results classifying data from IFG ROI (Go-NoGo task) with the same features, their classification reach 70% of the accuracy as maximum. Thus, we can conclude that the machine learning approach performs a good classification between groups using fMRI data from both ROIs during the tasks.

The best features, comparing the betas to the contrasts, are betas in both cases and without a previous feature selection, since worse results are obtained using main betas. Given that the accuracy obtained with contrasts features is always lower than the accuracy from the betas features, we can say the best features to perform an accurate classifier are beta parameters. This is probably due to the fact that beta parameters contains the effect size (activation level of the ROI) from specific conditions of the experiment (for example, when the subject is waiting a go trial signal, or when there is the possibility to win money) and it is more discriminative between groups, rather than the contrasts, where the bigger discrimination is between different conditions in the same task.

In case of fitted and EOI fitted data, we do not obtain an accurate classifier for the MID task. Otherwise these temporal features perform a good work for IFG data (70% of accuracy), using feature dimensionality reduction in both cases.

Analyzing the feature dimensionality reduction, in betas and contrasts parameters, it is clearly seen that it does not improve the classification, neither if it is a PCA nor a feature selection (main betas). In contrast, in fitted and EOI fitted data, due to the large number of dimensions (the feature is a time point, so it is 180 per side ROI), a dimensionality reduction is useful. Thus, the feature extraction (left ROI) in fitted data and the feature extraction (PCA) in EOI improve the results.

From the point of view of the classifier selection, LinearSVM is the best model for classification in most of the cases, always choosing a very small cost of constraint violation parameter ($C < 1$). The parameters for which the RbfSVM model works better are small γ's (γ<0.1) and C values between 1 and 10.

Chapter 7

Conclusions and Future Work

In this project, we have explored fMRI data to classify ADHD subjects. We have reviewed the previous statistical study presented in [1] and we have validated their clinical hypotheses using new machine learning strategies. We have demonstrated that machine learning techniques in the analysis of fMRI data provides a valid tool for clinical classification.

In particular, our proposed approach is able to distinguish the ADHD characteristics in a subject based on the activation in the executive functions (inhibition response and reward anticipation). Comparing to results in [1], we have demonstrated that, by selecting discriminative features, group classification can be successfully performed. Moreover, opposite to the results on Go-NoGo task in [1], we have found accurate classification performance. We have also shown that classification rates can be significantly improved by incorporating temporal information into machine learning analysis.

To our knowledge, this is the first time that the temporal/functional information of the fMRI data is explicitly explored for machine learning classification purposes in ADHD patients.

Since feature selection is the key for pattern recognition problems, especially when only a small number of data are available, as in most human subject research, one of our future research directions will be to explore efficient global dimensionality reduction techniques proposed in the literature that can be applied on extremely high dimensional training data and examine more sophisticated classifiers. Another future work will be to apply the same group analysis approach in other diagnosed brain disorder.

Bibliography

[1] S. Carmona, E. Hoekzema, J. Ramos-Quiroga, V. Richarte, C. Canals, M. Bosch, R nd Rovira, J. Carlos Soliva, A. Bulbena, A. Tobeña, M. Casas, O. Vilarroya, Response inhibition and reward anticipation in medication-naïve adults with attention-deficit/hyperactivity disorder: A within-subject case-control neuroimaging study., Human Brain Mappingdoi:i: 10.1002/hbm.21368.

[2] Buxton, Rick, Introduction to Functional Magnetic Resonance Imaging (2002) .

[3] K. Friston, J. Ashburner, S. Kiebel, T. Nichols, W. Penny, Statistical Parametric Mapping: The Analysis of Functional Brain Images (2007) .

[4] T. Mitchell, Machine Learning, McGraw Hill, NY (1997) .

[5] M. Masaya, K. You, B. Peter A., K. Nikolaus, Comparison of multivariate classifiers and response normalizations for pattern-information fMRI, Neuroimage 53 (2010) 103–118.

[6] L. Zhang, D. Samaras, D. Tomasi, N. Alia-Klein, L. Cottone, A. Leskovjan, N. Volkow, R. Goldstein, Exploiting temporal information in functional magnetic resonance imaging brain data, Proc. of the Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv (2005) 679–687.

[7] G. Polanczyk, M. Lima, B. Horta, J. Biederman, L. Rohde, The worldwide prevalence of ADHD: A systematic review and metaregression analysis, Am J Psychiatry 164 (2007) 942–948.

[8] J. Biederman, C. Petty, A. Clarke, A. Lomedico, S. Faraone, Predictors of persistent ADHD: An 11-year follow-up study, Psychiatr Res J 45 (2011) 150–155.

[9] A. Aron, R. Poldrack, The cognitive neuroscience of response inhibition: Relevance for genetic research in attention- deficit/hyperactivity disorder, Biol Psychiatry 57 (2005) 1285–1292.

[10] M. Plichta, N. Vasic, R. Wolf, K. Lesch, D. Brummer, C. Jacob, G. Fallgatter, AJ amd Gron, Neural hyporesponsiveness and hyperresponsiveness during immediate and delayed reward processing in adult attention-deficit/hyperactivity disorder, Biol Psychiatry 65 (2009) 7–14.

[11] A. Scheres, M. Milham, B. Knutson, F. Castellanos, Ventral striatal hyporesponsiveness during reward anticipation in attention-deficit/hyperactivity disorder., Biol Psychiatry 61 (2007) 720–724.

[12] A. Strohle, M. Stoy, J. Wrase, S. Schwarzer, F. Schlagenhauf, M. Huss, J. Hein, A. Nedderhut, B. Neumann, A. Gregor, G. Juckel, B. Knutson, U. Lehmkuhl, M. Bauer, A. Heinz, Reward anticipation and outcomes in adult males with attention-deficit/ hyperactivity disorder., Neuroimage 39 (2008) 966–972.

[13] S. Chamberlain, A. Hampshire, R. K. Muller, U and, C. Del, K. Craig, R. Regenthal, J. Suckling, J. Roiser, B. E. Grant, JE and, T. Robbins, B. Sahakian, Atomoxetine modulates right inferior frontal activation during inhibitory control: A pharmacological functional magnetic resonance imaging study, Biol Psychiatry 65 (2009) 550–555.

[14] P. Shaw, W. Sharp, M. Morrison, K. Eckstrand, D. Greenstein, L. Clasen, A. Evans, J. Rapoport, Psychostimulant treatment and the developing cortex in attention deficit hyperactivity disorder, Am J Psychiatry 166 (2009) 58–63.

[15] J. Ashburner, G. Barnes, C.-C. Chen, J. Daunizeau, G. Flandin, K. Friston, S. Kiebel, J. Kilner, V. Litvak, R. Moran, W. Penny, M. Rosa, K. Stephan, D. Gitelman, R. Henson, C. Hutton, V. Glauche, J. Mattout, C. Phillips, SPM8 Manual (2011) .

[16] A. J. Bartsch, G. Homola, A. Biller, L. Solymosi, M. Bendszus.

[17] P. Bandettini, A. Jesmanowicz, E. Wong, J. Hyde, Processing strategies for time-course data sets in functional MRI of the human brain, Magn. Reson. Med. 30 (1993) 161–173.

[18] K. Friston, A. Holmes, K. Worsley, J. Poline, C. Frith, R. Frackowiak, Statistical parametric maps in functional imaging: a general linear approach, Hum. Brain Mapp. 2 (1995) 189–210.

[19] K. Worsley, K. Friston, Analysis of fMRI time-series revisited-again, NeuroImage 2 (1995) 173–181.

[20] Z. Wang, C. A. R., J. Wang, J. A. Detrea, Support vector machine learning-based fMRI data group analysis, NeuroImage 36 (2007) 1139–1151.

[21] K. Friston, P. Jezzard, R. Turner, Analysis of functional MRI timeseries, Hum. Brain Mapp. 1 (1994) 153–171.

[22] J. Hopfinger, C. Buchel, A. Holmes, K. Friston, A study of analysis parameters that influence the sensitivity of event-related fMRI analyses, NeuroImage 11 (2000) 326–333.

[23] Z. Wang, A hybrid SVM?GLM approach for fMRI data analysis, NeuroImage 46 (2009) 608–615.

[24] G. Aguirre, E. Zarahn, M. D'Esposito, The variability of human BOLD hemodynamic responses, NeuroImage 8 (1998) 360–369.

[25] D. Cox, R. Savoy, Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex, NeuroImage 19 (2003) 261–270.

[26] C. Davatzikos, K. Ruparel, Y. Fan, D. Shen, M. Acharyya, J. Loughead, R. Gur, D. Langleben, Classifying spatial patterns of brain activity with machine learning methods: application to lie detection, NeuroImage 28 (2005) 663–668.

[27] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereira, X. Wang, M. Just, S. Newman, Learning to decode cognitive states from brain images, Mach. Learning 57 (2004) 145–175.

[28] L. Zhang, D. Samaras, V. N. G. R. Tomasi, D., Machine learning for clinical diagnosis from functional magnetic resonance imaging, IEEE International Conference Computer Vision and Pattern Recognition (2005) 1211–1217.

[29] S. LaConte, S. Strother, V. Cherkassky, J. Anderson, X. Hu, Support vector machines for temporal classification of block design fMRI data, NeuroImage 26 (2005) 317–329.

[30] J. Mourão-Miranda, A. Bokde, C. Born, H. Hampel, M. Stetter, Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data, NeuroImage 28 (2005) 980–995.

[31] K. Friston, C. Büchel, Functional connectivity: eigenimages and multivariate analysis, Human Brain Function (1997) .

[32] M. McKeown, S. Makeig, G. Brown, T. Jung, S. Kindermann, A. Bell, T. Sejnowski, Analysis of fMRI data by blind separation into independent spatial components, Hum. Brain Mapp. 6 (1998) 160–188.

[33] A. Bell, T. Sejnowski, An information maximisation approach to blind separation and blind deconvolution, Neural Comput 7 (1995) 1129–1159.

[34] X. Wang, R. Hutchinson, T. Mitchell, Training fMRI classifiers to discriminate cognitive states across multiple subjects, The 17th Annual Conference on Neural Information Processing Systems (2003) .

[35] J. Ford, H. Farid, F. Makedon, L. Flashman, T. McAllister, V. Megalooikonomou, A. Saykin, Patient classification from fMRI activation maps. 6th Annual International Conference on Medical Image Computing and Computer Assisted Intervention (2003) .

[36] N. Kriegeskorte, R. Goebel, P. Bandettini, Information-based functional brain mapping, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 3863–3868.

[37] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, E. Formisano, Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns, NeuroImage 43(1) (2008) 44–58.

[38] M. Björnsdotter Äberg, L. Löken, J. Wessberg, An evolutionary approach to multivariate feature selection for fMRI pattern analysis, Biosignals 2 (2008) 302–307.

[39] D. B. Rowe, R. G. Hoffman, Multivariate Statistical Analysis in fMRI, IEEE engineering in medicine and biology magazine.

[40] U. Kjems, L. Hansen, J. Anderson, S. Frutiger, S. Muley, J. Sidtis, D. Rottenberg, S. Strother, The quantitative evaluation of functional neuroimaging experiments: mutual information learning curves, NeuroImage 15 (2002) 772–786.

[41] S. LaConte, J. Anderson, S. Muley, J. Ashe, S. Frutiger, K. Rehm, L. Hansen, E. Yacoub, X. Hu, D. Rottenberg, S. Strother, The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics., NeuroImage 18 (2003) 10–27.

[42] T. Carlson, P. Schrater, S. He, Patterns of activity in the categorical representations of objects, J. Cogn. Neurosci 15 (2003) 704–717.

[43] S. Strother, S. La Conte, L. Kai Hansen, J. Anderson, J. Zhang, S. Pulapura, D. Rottenberg, Optimizing the fMRI dataprocessing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis, NeuroImage 23 (2004) S196–S207.

[44] J. Haxby, M. Gobbini, M. Furey, A. Ishai, J. Schouten, P. Pietrini, Distributed and overlapping representations of faces and objects in ventral temporal cortex, Science 293 (2001) 2425–2430.

[45] E. Formisano, F. De Martino, G. Valente, Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning, Science Direct 26 (2008) 921–934.

[46] F. Pereira, T. Mitchell, M. Botvinick, Machine learning classifiers and fMRI: a tutorial overview., NeuroImage 45 (1 Suppl) (2009) S199–S209. doi:10.1016/j.neuroimage.2008.11.007.
URL `http://dx.doi.org/10.1016/j.neuroimage.2008.11.007`

[47] S. Carmona, E. Proal, E. A. Hoekzema, J.-D. D. Gispert, M. Picado, I. Moreno, J. C. C. Soliva, A. Bielsa, M. Rovira, J. Hilferty, A. Bulbena, M. Casas, A. Tobeña, O. Vilarroya, Ventro-striatal reductions underpin symptoms of hyperactivity and impulsivity in attention-deficit/hyperactivity disorder., Biological psychiatry 66 (10) (2009) 972–977. doi:10.1016/j.biopsych.2009.05.013.

[48] S. Durston, T. NT, K. Thomas, M. Davidson, I. Eigsti, Y. Yang, A. Ulug, B. Casey, Differential patterns of striatal activation in young children with and without ADHD. Biol Psychiatry, Biol Psychiatry 53 (2003) 871–878.

[49] L. Tamm, V. Menon, J. Ringel, A. Reiss, Event-related FMRI evidence of frontotemporal involvement in aberrant response inhibition and task switching in attention-deficit/ hyperactivity disorder, J Am Acad Child Adolesc Psychiatry 43 (2004) 1430–1440.