Miguel Claudino Leão Garrett Fernandes

Bachelor Degree in Biomedical Engineering Sciences

Assessment of a multi-measure functional connectivity approach

Dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Science in
Biomedical Engineering

Adviser: Alexandre Andrade, Assistant Professor,
Institute of Biophysics and Biomedical Engineering of the University of Lisbon

Co-adviser: Ricardo Vigário, Assistant Professor,
Faculty of Sciences and Technology of the Nova University of Lisbon

Examination Committee

Chairperson: Carla Quintão Pereira
Raporteur: Maria Margarida Silveira
Member: Alexandre Andrade

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

December, 2017

Assessment of a multi-measure functional connectivity approach

This document was created using the (pdf)LaTeX processor, based in the "novathesis" template[1], developed at the Dep. Informática of FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Acknowledgements

I wish to thank to my adviser, professor Alexandre Andrade, and to PhD Susana Santos for all their guidance, availability and kindness throughout the whole process of writing this thesis. It was a pleasure working with both of you. To PhD Pedro Guimarães for his availability, to professor Fernando Batista of ISCTE also for his availability and especially for his interest and enthusiasm, to professor Branislav Gerazov for his brief but invaluable help in my introduction to machine learning. Also, to everyone I had the opportunity to meet at the institute and that made me feel welcomed there. Specially, to MSc. Diogo Duarte for always sharing his wise thoughts on all my questions and to his organization of many journal club meetings at the institute, to which I had the pleasure of attending.

Because this work really is the product of five years of education, I would like to thank professor Ricardo Vigário and to all my professors at the Faculty of Sciences and Technology of the Nova University of Lisbon and specially to those who cultivated my passion for science and for teaching with their enthusiasm, knowledge and patience. Also, for the same reasons, to the ones at the Polytechnic of Milan during my Erasmus program in Italy.

To my family for their support, for their belief in me and for allowing me to pursue my life goals, thank you. Shall we keep playing many more chess games, grandfather! (and shall you keep winning them too).

To all my friends for making this five-year journey a much better one and specially to Marta Pinto for her help in this thesis and for everything else, but mostly for her invaluable friendship. I hope you all keep making part of my life in some way or another for years to come.

Finally, to the greats of science of the past and present for inspiring me through their work and creativity and for their lessons of resilience and humility.

Look again at that dot. That's here. That's home. That's us. On it everyone you love, everyone you know, everyone you ever heard of, every human being who ever was, lived out their lives. The aggregate of our joy and suffering, thousands of confident religions, ideologies, and economic doctrines, every hunter and forager, every hero and coward, every creator and destroyer of civilization, every king and peasant, every young couple in love, every mother and father, hopeful child, inventor and explorer, every teacher of morals, every corrupt politician, every "superstar," every "supreme leader," every saint and sinner in the history of our species lived there - on a mote of dust suspended in a sunbeam.

Carl Sagan, in: Pale Blue Dot: A Vision of the Human Future in Space, about a photograph of the planet Earth taken by the Voyager 1 from a distance of 6×10^9 *km.*

Abstract

Efforts to find differences in brain activity patterns of subjects with neurological and psychiatric disorders that could help in their diagnosis and prognosis have been increasing in recent years and promise to revolutionise clinical practice and our understanding of such illnesses in the future. Resting-state functional magnetic resonance imaging (rs-fMRI) data has been increasingly used to evaluate said activity and to characterize the connectivity between distinct brain regions, commonly organized in functional connectivity (FC) matrices. Here, machine learning methods were used to assess the extent to which multiple FC matrices, each determined with a different statistical method, could change classification performance relative to when only one matrix is used, as is common practice. Used statistical methods include correlation, coherence, mutual information, transfer entropy and non-linear correlation, as implemented in the MULAN toolbox. Classification was made using random forests and support vector machine (SVM) classifiers. Besides the previously mentioned objective, this study had three other goals: to individually investigate which of these statistical methods yielded better classification performances, to confirm the importance of the blood-oxygen-level-dependent (BOLD) signal in the frequency range 0.009-0.08 Hz for FC based classifications as well as to assess the impact of feature selection in SVM classifiers. Publicly available rs-fMRI data from the Addiction Connectome Preprocessed Initiative (ACPI) and the ADHD-200 databases was used to perform classification of controls vs subjects with Attention-Deficit/Hyperactivity Disorder (ADHD). Maximum accuracy and macro-averaged f-measure values of 0.744 and 0.677 were respectively achieved in the ACPI dataset and of 0.678 and 0.648 in the ADHD-200 dataset. Results show that combining matrices could significantly improve classification accuracy and macro-averaged f-measure if feature selection is made. Also, the results of this study suggest that mutual information methods might play an important role in FC based classifications, at least when classifying subjects with ADHD.

Keywords: fMRI, classification, functional connectivity matrices, SVM, feature selection, mutual information

Resumo

A identificação de diferenças nos padrões de atividade cerebral de indivíduos com doenças mentais poderá revolucionar o conhecimento relativamente às causas subjacentes, assim como a capacidade de diagnóstico e prognóstico em contexto clínico. A imagem por ressonância magnética funcional em repouso (rs-fMRI) tem sido largamente utilizada para avaliar esta atividade e caracterizar a conetividade entre diferentes regiões cerebrais. Esta informação é normalmente organizada em matrizes de conetividade funcional (FC). Neste estudo, utilizaram-se técnicas de aprendizagem automática para avaliar de que forma diferentes matrizes de conetividade, usadas em simultâneo, alterariam a qualidade de uma classificação automática relativamente ao caso comum em que apenas uma matriz é utilizada. Os métodos estatísticos calculados com o programa MULAN incluem: correlação, coerência, informação mútua, transferência de entropia e ainda correlação não linear. A classificação fez-se com recurso a *random forests* e a *support vector machines* (SVMs). Adicionalmente, três outros objetivos foram traçados: comparar a qualidade da classificação obtida com recurso a cada método individualmente, confirmar a importância da informação do sinal *blood-oxygen-level-dependent* (BOLD) na gama de frequências 0.009-0.08 Hz para a classificação automática baseada em FC e finalmente avaliar o impacto da seleção de atributos em classificadores SVM. Utilizaram-se dados de rs-fMRI de duas bases de dados públicas - *Addiction Connectome Preprocessed Initiative* (ACPI) e ADHD-200 - para classificação de sujeitos de controlo vs sujeitos com Transtorno de Défice de Atenção e Hiperatividade (TDAH). Os valores máximos de precisão e *macro-averaged f-measure* foram, respetivamente, 0.744 e 0.677 no conjunto de dados da ACPI e, nos da ADHD-200, de 0.678 e 0.648. Os resultados obtidos mostram que a combinação de matrizes pode aumentar significativamente a qualidade da classificação se existir seleção prévia de atributos. Mais, este estudo sugere que a informação mútua poderá desempenhar um papel importante na classificação automática baseada em FC de sujeitos com TDAH.

Palavras-chave: fMRI, classificação automática, matrizes de conetividade funcional, SVM, selecção de atributos, informação mútua

Contents

List of Figures

List of Tables

Acronyms

AAL	Automated Anatomical Labeling.
ACPI	Addiction Connectome Preprocessed Initiative.
AD	Alzheimer's Disease.
ADC	Apparent Diffusion Coefficient.
ADHD	Attention-Deficit/Hyperactivity Disorder.
ANN	Artificial Neural Network.
ANOVA	Analysis of Variance.
ANTs	Advanced Normalization Tools.
AUC	Area Under the Curve.
BOLD	Blood-Oxygen-Level-Dependent.
C-PAC	Configurable Pipeline for the Analysis of Connectomes.
DCM	Dynamic Causal Modelling.
DL	Deep Learning.
DTI	Diffusion Tensor Imaging.
EC	Effective Connectivity.
EEG	Electroencephalography.
EPI	Echo Planar Imaging.
FC	Functional Connectivity.
FCC	Fully Connected Cascade.
FCP	1000 Functional Connectomes Project.
FID	Free Induction Decay.
fMRI	Functional Magnetic Resonance Imaging.
FN	False Negatives.
FP	False Positives.

FPR	False Positive Rate.
HASTE	Half-Fourier Acquisition Single-shot Turbo Spin Echo.
HC	Healthy Controls.
INDI	International Neuroimaging Data-Sharing Initiative.
K-NN	K-Nearest Neighbours.
LASSO	Least Absolute Shrinkage and Selection Operator.
LOO	Leave-One-Out.
MDD	Major Depressive Disorder.
MEG	Magnetoencephalography.
MFC	Most Frequent Class.
ML	Machine Learning.
MNI	Montreal Neurological Institute.
MRI	Magnetic Resonance Imaging.
MRMD	Max Relevance Max Distance.
MTA 1	Multimodal Treatment of Attention Deficit Hyperactivity Disorder 1.
MULAN	Multiple Connectivity Analysis.
MVAR	Multivariate Autoregressive.
OCD	Obsessive-compulsive Disorder.
PCA	Principal Component Analysis.
PCP	Preprocessed Connectomes Project.
PET	Positron Emission Tomography.
RBF	Radial Basis Function.
RF	Random Forest.
ROC	Receiver Operating Characteristics.
ROI	Region Of Interest.
ROIs	Regions Of Interest.
rs-fMRI	resting-state fMRI.
SNR	Signal to Noise Ratio.

SPECT	Single-Photon Emission Computer Tomography.
SVM	Support Vector Machine.
TN	True Negatives.
TNR	True Negative Rate.
TP	True Positives.
TPR	True Positive Rate.

Chapter 1

Introduction

Since the mid 19th century, when Dr. John M. Harlow reported the now famous case of a man, Phineas Gage, that dramatically changed personality after surviving the destruction of his left frontal lobe provoked by an accident with an iron bar [65], that the human brain has been indubitably associated with cognitive function. Our knowledge of the brain and of its physiology has greatly increased since then, fast-forward one century and not only non-invasive techniques to anatomically image the brain and record its electrical activity had been developed but also techniques to image its functioning and activity over time. Despite that, years after the development of such techniques, we are still far from conquering the inherent complexity of said organ. We clearly can gather knowledge about the world around us but what we still do not know is how we/our brains do it.

Not too long after brain activity started being commonly recorded, it was hypothesised that if the activity from two neurons had a significant deviation from statistical independence, then, it could be that they were connected in some way [108]. This connection was then called Functional Connectivity (FC) [5]. This idea was generalized for populations of neurons and, in the early 1990s, the hypothesis that two brain regions could be functionally connected in individuals suffering from neurological and psychiatric disorders and in healthy subjects in different ways was already established (e.g. [87]). In the same decade, the development of Functional Magnetic Resonance Imaging (fMRI) [13, 104], which measures brain activity using Blood-Oxygen-Level-Dependent (BOLD) contrast, greatly boosted research in this field and the FC between several brain regions of different classes of subjects during a given task started to be explored. In the last decade, it was proposed that brain activity acquired while subjects are at rest (on a resting state) could potentially give helpful information about their FC [62]. Since then, functional imaging techniques, including fMRI started being acquired while patients were at rest with the purpose of assessing their whole-brain FC patterns. These patterns are usually

stored as matrices, called FC matrices.

Mental disorders are still currently defined and diagnosed mostly based on the inter-subject consistency of a set of personality traits and behavioural symptoms [10]. This diffuse characteristic of psychiatry is not optimal and one would like to establish a more direct cause-disorder relationship. Thus, if different FC patterns are actually in the root of some mental disorders, the identification of such patterns could potentially provide new ways to define such disorders and greatly improve the diagnostic and prognostic tools of neurologists, psychiatrists and psychologists.

How would one identify group-differences in whole-brain connectivity patterns? That is where Machine Learning (ML) techniques enter into scene. Automatic pattern recognition techniques have been developed long since and tasks such as this one are exactly the reason for which they are developed. Besides recognizing whether differences in FC patterns between groups of subjects actually exist, these techniques provide rules that can be used to classify a subject as belonging to one group or the other, in case the rule is true.

Several mathematical methods have been used to quantify the FC between two distinct brain regions, including correlation, the first one to be used, and coherence. Usually, FC patterns are derived using only one mathematical method to calculate the FC between a given brain region and all others. *In this work, connectivity patterns are going to be derived using several mathematical methods simultaneously, and the separability of two classes using these patterns is going to be compared to the separability of the two same classes using patterns derived with only one mathematical method.*

More specifically this study has four goals:

1. The first and main goal is to compare the classification performance of an ML classifier when using only one FC matrix to extract features, with its performance in the same conditions but using several FC matrices, each derived using a different mathematical method, to extract features;

2. To investigate which statistical method derives the FC matrix that yields the best classification performances;

3. To evaluate the impact of feature selection in FC based classifications;

4. To confirm the importance of low BOLD signal oscillations (< 0.1 Hz) in such classifications.

In order to achieve these goals, brain activity data from two classes of subjects was needed to build the FC matrices to further use them to extract features for the previously mentioned classifications. Publicly available resting-state fMRI (rs-fMRI) data of subjects with Attention-Deficit/Hyperactivity Disorder (ADHD) and controls from the ACPI and the ADHD-200 databases were used with this purpose. Classification of subjects with ADHD has not been very successful, specially when compared to what has already been

achieved for other neurological and psychiatric disorders [153], which emphasizes the need for better ADHD classification methodologies than the ones currently in practice. Also, each mathematical method highlights different relationships among brain activity signals, which means that significantly better classification performances using a specific method or a combination of them, could give insight on which relationships between different regions actually reflect the biological causes behind the mental disorder in analysis.

The remaining of this text is structured as follows:

Chapter 2 is intended to introduce the reader to the main concepts used in the following text, while developing them to an extent that could ideally allow him/her to derive their own conclusions from the reported results.

Chapter 3 provides the state-of-art in classification of subjects with ADHD, specifically in the two databases used in this study, as well as the best available answers to the questions this study is intended to approach.

Chapter 4 thoroughly explains the methodology used to achieve the previously mentioned goals in such a manner that the interested reader could closely replicate this study.

Chapter 5 briefly analyses the separability of the two classes in order to evaluate the feasibility of the classifications and if the classification pipeline works properly.

Chapter 6 reports the results achieved using the methodology described in Chapter 4 and their analysis.

Chapter 7 discusses the achieved results, compares them with what was hypothesised using the knowledge from Chapters 2 and 3 and provides the author's thoughts on the results of those comparisons.

Chapter 8 provides a summary of the conclusions regarding the goals of this work and, along with Chapter 7, describes possible future research directions.

Chapter 2

Theoretical Background

The present Chapter is intended to introduce the reader to key concepts needed to understand the following text. By the end of the Chapter, it should be clear how one can try to automatically distinguish two or more groups of people with different neurological characteristics using brain activity. The reader is expected to be more or less acquainted with Magnetic Resonance Imaging (MRI), one of the most important techniques used to inspect the brain through bone and tissue and the one that originated the data used in this work. Notwithstanding, a very short introduction to MRI is going to be made next and, the remaining of the Chapter will be built on that.

2.1 Imaging the Brain

2.1.1 Magnetic Resonance

Let us consider a sample of an element in a constant magnetic field , $\vec{B_0}$, with a given intensity and direction. If the nucleus of that element has non-zero spin, then it is going to acquire a precession movement around a rotation axis with the same direction of $\vec{B_0}$, as a consequence of the torque applied by the magnetic field [128, pp. 209-211]. Additionally, in that case, the nuclear magnetic dipole moment of the element's atoms might be oriented, relative to $\vec{B_0}$, in a finite number of ways, depending on the spin magnetic moment of the nucleus [80, p. 398]. Each of these orientations corresponds to a given potential energy and, at low enough temperatures, the privileged one is that with the lowest energy. So, when a non-zero spin nucleus is put in an external magnetic field, its magnetic dipole moment precesses around the magnetic field and its energy varies depending on the acquired orientation. This energy split is what is commonly referred to as the Zeeman effect [80, p. 398].

In the human body, the predominant element with non-zero nuclear spin is the ^{1}H

(both ^{12}C and ^{16}O nuclei have null spin [30]) such that, to a first approximation, one can consider only the effect of the ^{1}H nuclei, when analysing the effect of an external magnetic field on the body [24, p. 3]. The component of the magnetic dipole moment of a proton, the nucleus of a ^{1}H atom, codirectional to an external field similar to $\vec{B_0}$, can be either $\pm 1/2$ in $\hbar$ units, which means that they can only have two orientations and, correspondingly, two energy levels in that case [128, pp. 206-208]. The distribution of protons between these two energy levels depends on the thermal energy and on the intensity of the external magnetic field [84, pp. 284-300]. The energy gap, ΔE, between the two levels, corresponds to the energy of a photon with a frequency equal to the magnetic dipole moment precession frequency, commonly referred to as Larmor frequency or ω_0 [128, pp. 211-212]. When an alternate magnetic field with frequency ω_0, varying perpendicularly to $\vec{B_0}$, is put over the precessing protons, these start to precess in phase with each other and with the field. Also, at the same time, some protons in the lowest energy level absorb an energy equal to the energy gap and pass to the highest energy level, changing their orientation to one regularly called "anti-parallel" to the constant external magnetic field (as opposed to the lowest energy level orientation, parallel to the same field)[128, p. 212]. This phenomenon of energy absorption when the two frequencies, the one from the alternate external magnetic field and the one from the precession movement, are equal, is called magnetic resonance [112].

2.1.2 Magnetic Resonance Imaging

It was previously said that if one puts an alternate magnetic field perpendicular to $\vec{B_0}$ over the system constituted by the body plus the external constant magnetic field, some protons would change to the highest energy level. In the same way, when that alternate magnetic field is turned off, the distribution of protons between the two levels of energy, goes to a new equilibrium state. This new state might be similar to the one the system was in before the alternate field was turned on, if every other variable remained unchanged. Also, the precession phases of each proton will distribute randomly as before [24, pp. 8-9]. The variation in the magnetic field of the system caused by this proton relaxation, induces a current proportional to it. The intensity of the measured signal and its decay pattern reflect the type of tissue in analysis. For instance, if a given tissue has a higher proton density, the intensity of the measured signal is also going to be higher [128, p.214].

As explained, it is possible to distinguish different types of tissue using the phenomenon of magnetic resonance. The gap between that and imaging is spatial encoding. For instance, let us consider a phantom with fragments of different proton densities under a constant magnetic field. If an alternate magnetic field pulse, with frequency ω_0, is emitted such that there is magnetic resonance between the field and the nucleus, the relaxation pattern and the current intensity are going to be a result of the variation in the magnetic field caused by all fragments at the same time. In such cases, it is not possible to identify which point of space caused which part of the signal. One way of achieving

that is by using gradient magnetic fields [128, pp. 223]. Since the Larmor frequency is a function of the intensity of the magnetic field over the sample [24, p. 2], a gradient field introduces a spatial dependence to the equation. Because of the way the spatial encoding is made, the relaxation pattern (the Free Induction Decay (FID) signal) can be directly stored in the spatial frequencies space, or as it is also called in this case, the k-space. The information can, then, be converted to the image space using the inverse 2D Fourier transform [128, pp. 229-231].

Various alternate magnetic field pulses are usually needed for k-space to be completely covered, however, in detriment of some image resolution, faster acquisition techniques have been developed as the Half-Fourier Acquisition Single-shot Turbo Spin Echo (HASTE) imaging or the Echo Planar Imaging (EPI) [45].

2.1.3 Functional Magnetic Resonance Imaging

The variation of the magnetic resonance signal with the physicochemical properties of the medium and its short acquisition time, allow this type of medical image to be used for functional imaging. The BOLD technique is the most commonly used for fMRI and it is based on the dependence of the FID signal on the blood concentration of deoxyhemoglobin [104], that, due to it being a paramagnetic molecule, introduces a hypointense contrast in the MRI image [72, pp. 193].

The increase in metabolic activity in a given brain region, increases oxygen consumption *in loco*. This, in turn, increases the blood concentration of deoxyhemoglobin, leading to a rapid decrease of the magnetic resonance signal in the same area. The relative hypoxia in the activated region, originates a local increase in blood supply, decreasing the concentration of deoxyhemoglobin, causing a lasting increase of intensity in the MRI image [72, pp. 193-199]. Thus, while imaging the brain, one could associate its hyperintense areas with their previous activation. The variations of the BOLD signal are very small, which leads to a very low Signal to Noise Ratio (SNR). To overcome this, cyclic activation patterns are induced by cyclic stimulation patterns which enables researchers to average out part of the noise. Nonetheless, low SNRs are still one of the fundamental obstacles to the introduction of fMRI in clinical practice [49].

Brain activity at rest was sometimes considered as uninteresting information [48]. Recently, however, it has been brought to the spotlight, particularlly since the discovery of valuable activity patterns such as the Default Mode Network, which is linked to introspection [62]. To fMRI using BOLD contrast while the patient is at rest one usually calls rs-fMRI [49]. This type of fMRI, besides having a better SNR [49], has the benefit of being a passive medical exam with minimal patient collaboration [48]. Such a characteristic widens the target population to patients with physical or neurological disabilities that do not allow them to cooperate as needed for task-related fMRI exams.

The acquisition of both rs-fMRI and task-related fMRI is usually made using EPI for better temporal resolution, which, as said before, comes at the cost of some spatial

resolution [72, p. 202]. From the fMRI data and, particularly, from that acquired in a resting state, it is possible to evaluate the brain connectivity of a subject, something to be developed in the following Section of this Chapter.

2.2 Brain Connectivity

2.2.1 Structural Connectivity and the Connectome

The human brain is widely recognized as the most complex organ in our bodies and can surely be considered one of the most complex systems in the so far known Universe. Neurons are one of the structural units of the brain and the nervous system. They can receive, propagate and transmit stimuli, and together they form an ever changing neural network. The path a stimulus travels inside the neural network, as well as the biological action it triggers, depend on how that network is arranged i.e. on the neuronal pattern of connections, something defined as the *structural connectivity* of the brain, also called anatomical or neuroanatomical connectivity.

There is a concept associated with structural connectivity on which great expectations of it becoming an unprecedented tool to understand the processes behind cognition are being deposited. This concept is called the *connectome*. The connectome is a wiring map where every connection each neuron makes with the others is represented, with the goal of making a complete description of the structural connectivity of an organism [132]. Up to now, only one connectome has been completed, that of a nematode of the *C. elegans* species. There are, however, projects aiming at storing and sharing human structural connectivity data of that type [46, 133].

Currently, whole-brain structural connectivity of the human brain is understood mainly on the basis of the connections that white matter tracts establish between brain regions [131]. Because the diffusion of protons is anisotropic in such structures, these can be reconstructed using Diffusion Tensor Imaging (DTI) which is an MRI technique that uses as contrast the diffusion coefficient of protons in the tissues (or, in fact, the Apparent Diffusion Coefficient (ADC)) [96].

2.2.2 Functional and Effective Connectivity

Besides the physical connections between neurons or populations of neurons that directly give rise to structural connectivity, another concept linked with brain connectivity emerged from the analysis of its activation patterns. This concept, called *FC*, aims at measuring how different brain regions combine to perform a given task or thought, consciously or subconsciously. The definition of this concept is still not completely clear [127]. Despite that, the definitions available in literature all seem to agree on the statistical origin of the concept. In this work, FC is going to be considered as the statistical dependence between two neurological and time-dependent signals as defined by Friston in [54].

Another concept still not always objectively defined is that of *Effective Connectivity (EC)* [127]. The distinction and relationship between EC and FC has not always been clearly stated, which has led to some different interpretations of its definition within the neuroscience community. In an attempt to define these two concepts as coherently as possible, the definition of EC adopted here is also going to be based on the one given in the same paper by Friston. Having that into account, EC might be defined as the influence a population of neurons exerts over another, measured on a statistical level [54].

The most important distinction between FC and EC resides in the nature of the relationships between the neuronal populations they aim to quantify, being non-causal in the case of FC and causal in the case of EC (here causal relationship should be understood in the context of Wiener's definition [152]). Thus, measures of EC need to be directed i.e. the measure of EC between two regions X and Y needs to be able to attribute different values to the relationship $X \rightarrow Y$ and $Y \rightarrow X$. There is a multitude of methods available to measure the statistical dependence between neurological signals, some to be introduced in the next Section, and, for some statistical metrics, EC might be derived from FC. The signals used to calculate both connectivity types might come from Electroencephalography (EEG), Magnetoencephalography (MEG), fMRI or Positron Emission Tomography (PET) data. Due to their better temporal resolution, EEG and MEG are the preferential brain activity recording modalities for the determination of FC and EC.

2.2.3 Statistical Methods to Evaluate Functional and Effective Connectivity

The scientific field responsible for the study of statistical dependences between two different signals, having or not into account the influence of others, is multivariate statistics. Up to now, several methods from multivariate statistics have been applied to calculate FC and EC, each with its own origins and mathematical formulations. Taxonomically, these metrics can be divided into linear and non-linear, model based and model free, directed or non directed, based on time-domain representations or based on frequency-domain representations.

Cross-correlation and Pearson Correlation

Correlation is the most classical way of measuring the dependence between two neurological signals. Even today, the overwhelming majority of published studies in neuroscience use it as a measure of FC [54] so much so that even the definition of FC of some authors rests on the concept of correlation (e.g. [47]).

Let x_n and y_n be two random variables with $n = 1, ..., N$. The *cross-correlation* function between x_n and y_n is defined as:

$$R_{xy}(\tau) = E\left[x_n y_{n+\tau}\right] = \frac{1}{N-\tau} \sum_{k=1}^{N-\tau} x_n y_{n+\tau} \tag{2.1}$$

This function measures how linearly dependent the two variables are. If the maximum value happens for a given τ then one might establish a causal relationship between the two, even though this is not true for all cases. If the two variables are centred to zero and standardised by subtracting their respective averages and dividing next by their respective standard deviations, such that both x_n and y_n have zero average and unit variance, then R_{xy} ranges between -1 (when one varies linearly with the symmetric of the other) and 1 (when one varies linearly with the other), being positive when the two variables have a direct dependence and negative when they have an inverse dependence. If R_{xy} is zero, the two variables are linearly independent. By centring and standardising the variables, the cross-correlation function actually becomes a standardised covariance function. In this context, the *Pearson product-moment correlation coefficient*, r, is defined as the value of the centred and standardised cross-correlation function at zero lag ($\tau = 0$):

$$r = \frac{\sum_N (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_N (x_n - \bar{x})^2 \sum_N (y_n - \bar{y})^2}} \tag{2.2}$$

As shown by Rodgers and Nicewander in their paper from 1988, the Pearson coefficient can also be geometrically seen as the slope of the regression line between the two standardised variables [117].

Fourier Based and Wavelet Based Coherence

N. Wiener and A. I. Khinchine demostrated that the Fourier transform of the auto- and cross-correlation functions between two stationary random processes, $\{x_k(t)\}$ and $\{y_k(t)\}$, where k is the index of the sample space, gives their auto- and cross-spectral density functions, respectively, if the auto- and cross-correlations functions exist and their integral is finite [17, p. 199]. The relationships between these functions are, thus, also known as the Wiener-Khinchine relations. Mathematically, the spectral density functions can be expressed as:

$$S_{xx}(\omega) = \int_{-\infty}^{\infty} R_{xx}(\tau) e^{-i2\pi\omega\tau} \mathrm{d}\tau \tag{2.3}$$

$$S_{yy}(\omega) = \int_{-\infty}^{\infty} R_{yy}(\tau) e^{-i2\pi\omega\tau} \mathrm{d}\tau \tag{2.4}$$

$$S_{xy}(\omega) = \int_{-\infty}^{\infty} R_{xy}(\tau) e^{-i2\pi\omega\tau} \mathrm{d}\tau \tag{2.5}$$

The *Coherence* or *Magnitude Squared Coherence* function can be defined as the normalized magnitude of the cross-spectral density function. Mathematically it can be written as:

$$coh_{xy}(\omega) = \frac{|S_{xy}(\omega)|}{\left[S_{xx}(\omega) \cdot S_{yy}(\omega)\right]^{1/2}} \tag{2.6}$$

From equation 2.6 one can conclude that the coherence is a real-valued metric, function of the frequency, ω. Coherence measures the linear dependency between two random processes for a given frequency, and ranges between 0 and 1, being zero for two linearly independent random processes and 1 in the opposite case. In spite of being real-valued, coherence is sensitive to both phase and magnitude [120], since its definition involves delaying the signals in respect to each other.

For finite length samples of two random variables, or for two different signals, one can only estimate the spectral densities needed for the coherence expression. One way of doing this is by using the discrete Fourier (series) coefficients of the whole time window. This, however, does not yield a good estimate of the variables' spectral densities [14]. One technique used to improve this estimation is the Welch's method, which consists of dividing the whole time series in various temporal windows and then averaging their spectral density estimations. This, of course, only makes sense if the signals are stationary or close to being stationary because only in that case can we extrapolate the spectral characteristics of an entire signal from a given sub-sample of it.

In the particular case of event-related data, one can try to derive a better estimation by averaging the spectral densities over trials. Another alternative is to use a parametric approach, where signals are considered, for instance, autoregressive processes [79]. Parametric approaches usually give good estimations but are computationally expensive, specially when compared to non-parametric methods such as the ones described above.

Because stationarity is not common in neural signals [83], an estimation of coherence over time should be considered. To achieve that, one can use wavelets to determine temporal estimations of spectral densities.

A wavelet is a function, usually represented by the symbol Ψ, which satisfies certain mathematical criteria, such as having finite energy [3, p. 7]. Wavelets are localised in frequency and time and are used to transform signals to a representation dependent on these variables. There are several functions commonly used as wavelets but here the focus will be put on the complex Morlet wavelet, the most common one. Following [83], the Morlet wavelet can be defined for time, τ, and frequency, f, as:

$$\psi_{\tau,f}(u) = \sqrt{f} e^{i2\pi f(u-\tau)} e^{-\frac{(u-\tau)^2}{\sigma^2}} \tag{2.7}$$

where σ is the standard deviation of the Gaussian function defined by the third factor in equation 2.7, inversely proportional to the frequency, f. The Morlet wavelet can be seen as a complex sinusoid in a Guassian envelope. Both translation and dilation of the wavelet are possible by changing the values of τ and σ respectively. The continuous wavelet transform of a signal, x, can be defined for any wavelet function, Ψ_x, as:

$$W_x(\tau, f) = \int_{-\infty}^{\infty} x(u) \Psi_{\tau,f}{}^*(u) \mathrm{d}u \tag{2.8}$$

where * denotes the complex conjugation. It should be noted that the translation and dilation variables were here included in the definition of the wavelet function which in

this case is equation 2.7. This, however, is not always adopted by authors, being usual for these variables to appear only in the definition of the wavelet transform (e.g. [63]). Using the continuous wavelet transform one can compute the wavelet cross-spectral density between two signals x and y, as:

$$SW_{xy}(t,f) = \int_{t-\delta/2}^{t+\delta/2} W_x(\tau,f) W_y{}^*(\tau,f) \mathrm{d}\tau \tag{2.9}$$

where δ is a scalar. Using the Morlet wavelet, the cross-spectrum at a given frequency, f, is calculated for every time, t, using information within the interval $[-\delta/2+t; \delta/2+t]$ with size $(-\delta/2+t)-(\delta/2+t)=\delta$. This interval defines, thus, the temporal resolution of the wavelet. Within each interval the cross spectrum is determined in a very similar fashion to the temporal averaging described before, using a Gaussian as windowing function (see: [83]). The major differences lay on the variable width of both the Gaussian function and the interval δ, the two inversely proportional to the frequency f. For higher frequencies δ and σ decrease and the temporal resolution increases, for lower frequencies the opposite is true. The wavelet auto-spectrum density is calculated as in equation 2.9 but considering the transform and the conjugate of the same signal.

From the auto- and cross-spectral densities one can determine the *wavelet coherence* between two signals x and y, which can be written as:

$$Wcoh(t,f) = \frac{|SW_{xy}(t,f)|}{\left[SW_{xx}(t,f) SW_{yy}(t,f)\right]^{1/2}} \tag{2.10}$$

Wavelet coherence also ranges between 0, when the two signals are linearly independent and 1, when the signals are linearly dependent. When the two signals are linearly independent, however, coherence can be different than 0. This happens because the estimation of the auto- and cross-spectral densities is not perfect and, because of that, neither is the estimation of coherence. When one intends to verify if two signals are linearly dependent, a statistical test should be used to compare, for instance, the obtained value of coherence with the distribution for two random independent signals. Similar statistical tests can be used when analysing the cross spectrum against a population specific background spectra, usually defined as the mean time-averaged wavelet power spectrum, as stated in [120].

Much more can be said about wavelet spectra and wavelet coherence. For more information about the subject the interested reader is referenced, for instance, to [3, 83, 139].

Mutual Information

Both correlation and coherence measure linear relationships between two signals. We shall now delve into some statistical methods that do not assume linear relationships between variables and that, in fact, do not assume any specific relationships between them.

One of those metrics, based on concepts from information theory, is mutual information [124]. To define mutual information one should first introduce the concept of entropy. Historically, the entropy measure used for mutual information is the Shannon entropy [124]. Let us consider the histogram of a random variable x with L bins. The probability, p_i^x of an occurrence of x to fall in a given bin i, $i = 1, 2, ..., L$, might be simply defined as $p_i^x = n_i/N$ where n_i is the previous number of occurrences that fell in i and $N = \sum_{i=1}^{L} n_i$ is the total number of occurrences. Following that, the Shannon entropy of x is given by [118]:

$$H(x) = -\sum_{i \in L^*} p_i^x \ln p_i^x \tag{2.11}$$

where L^* is the set of bins with $n_i \neq 0$.

The previous definition of p_i^x introduces an underestimation of entropy for a finite number of samples of x when $p_i^x = 0$ [38]. To compensate for this, a corrected form of entropy might be defined following [118] as:

$$H_\infty(x) = H(x) + \frac{|L^*| - 1}{2N} \tag{2.12}$$

where $|\cdot|$ is the cardinality of the set. The Shannon entropy can be seen as the expected value of $\ln p_i$. This leads to low entropies for "concentrated" distributions and high entropies for "spread" distributions. In fact, the entropy is maximum when the distribution of x is uniform and zero when all occurrences are in the same bin. This happens because, for values smaller than 1, the negative of the logarithmic function favours lower values and greatly penalizes high values, so, if the distribution is spread, there will be more values close to zero and the entropy will be higher.

As was done for a single random variable, one can also define the entropy between two random variables, x and y. Considering y segmented into M bins and p_j^y the probability of an occurrence of y to fall in j, the joint entropy between x and y is given by:

$$H(x, y) = -\sum_{i \in L^*} \sum_{j \in M^*} p_{ij}^{xy} \ln p_{ij}^{xy} \tag{2.13}$$

where $p_{ij}^{xy} = n_{ij}/N$ with n_{ij} being the number of occurrences of x in i, when y falls in j. Using the joint entropy one can define the *mutual information* between x and y as:

$$MI(x, y) = H(x) + H(y) - H(x, y) = \sum_{i \in L^*} \sum_{j \in M^*} p_{ij}^{xy} \ln \frac{p_{ij}^{xy}}{p_i^x p_j^y} \tag{2.14}$$

again, following [118], one correction to 2.14 is:

$$MI_\infty(x, y) = MI(x, y) + \frac{|L^*| + |M^*| - |LM^*| + 1}{2N} \tag{2.15}$$

where LM^* is the set of pairs of bins that satisfy $n_{ij} \neq 0$.

Equation 2.14 allows us to understand what mutual information is actually measuring. If x and y are independent, then $p_{ij}^{xy} = p_i^x p_j^y$ and $H(x,y) = H(x) + H(y)$, which results in $MI = 0$. Thus, mutual information is always greater or equal than zero and measures the difference between the joint entropy when one assumes the two variables as independent and the actual joint entropy. When the two variables are similar MI has its maximum value and if the two variables are commuted the result is the same ($MI(x,y) = MI(y,x)$ because $H(x,y) = H(y,x)$), which means that MI gives no information about causality. One way to introduce directionality (and hence possibly causality) in mutual information is to introduce a time lag in one of the variables and calculate equation 2.14 for different lags [144].

Transfer Entropy

Delayed mutual information does give us information about direction, but there is, however, another statistical method, also based on information theory, that was defined specifically to measure the flow of information. This measure, called transfer entropy, was introduced by Thomas Schreiber in 2000 [121]. Transfer entropy is based on the concepts of transition probability and entropy rate. If a time series x can be approximated by a stationary Markov process of order k then the time-conditional probability of finding x in the state (or the bin, if we are still looking at the variable's histogram) i_{n+1} at time $n+1$ is independent of the state i_{n-k}. The entropy rate of x, then, measures how much entropy x has, given the knowledge of its previous k states. Mathematically, the entropy rate of x can be expressed as:

$$H_{rate}(x) = -\sum_{i_{n+1},\ldots,i_{n-k+1}} p^x(i_{n+1},\ldots,i_{n-k+1}) \ln p^x(i_{n+1}|i_n,\ldots i_{n-k+1}) \tag{2.16}$$

Both mutual information and transfer entropy can be understood in the context of the Kullback entropy [82], which might be expressed, for a given variable, as $K = \sum_i p^x(i) \ln p^x(i)/q^x(i)$ where $q(i)$ is the assumed probability distribution and $p^x(i)$ the actual probability distribution. The Kullback entropy measures the error of assuming the wrong probability distribution $q^x(i)$. Mutual information can, thus, be seen as the joint Kullback entropy between two variables x and y, when assuming the two variables independent or, mathematically, assuming that $q^{xy}(i,j) = p^x(i)p^y(j)$. Transfer entropy shall be here defined too in the context of the Kullback entropy as Schreiber did in his paper. Let us consider the following Markov property:

$$p^x(i_{n+1}|i_n,\ldots i_{n-k+1}) = p^{xy}(i_{n+1}|i_n,\ldots,i_{n-k+1},j_n,\ldots,j_{n-l+1}) \tag{2.17}$$

The assumption of 2.17 when there actually is information flow from y to x, leads to an error that can be measured by a conditional Kullback entropy, which in this case we call *Transfer Entropy* [121]:

$$TE_{y\to x} = \sum_{i_{n+1},\ldots,i_{n-k+1},j_n,\ldots,j_{n-k+1}} p^{xy}(i_{n+1}, i_n^{(k)}, j_n^{(l)}) \ln \frac{p^{xy}(i_{n+1}|i_n^{(k)}, j_n^{(l)})}{p^{x}(i_{n+1}|i_n^{(k)})} \tag{2.18}$$

where i_n and j_n are, as before, the possible states of, respectively, x and y at time n. Also, a shorthand notation $i_n^{(k)} = (i_n, \ldots, i_{n-k+1})$ and $j_n^{(k)} = (j_n, \ldots, j_{n-k+1})$ was here used as introduced by Schreiber.

By measuring the deviation from property 2.17, transfer entropy attempts to measure the information theory equivalent of Wiener's causality. Thus, transfer entropy not only is a directed measure, but also a model-free candidate to measure EC. For more information on transfer entropy and its applications to neuroscience the reader is referred to [145, 151].

Non Linear Correlation Coefficient

In 1989 Lopes da Silva et al. [89] first used the non linear correlation coefficient to measure the relationship between EEG signals. The idea behind this coefficient is to consider a given dependent variable, y_n, of finite length N as a function of another variable, x_n, and calculate the fraction of the total variance of y_n that can be explained by x_n. This measure is called η^2 and to estimate it one can use the regression curve to estimate y_n from x_n and calculate the unexplained variance of y_n using the estimation errors. The explained variance of y_n is then calculated by subtracting the unexplained variance from the total variance. The estimation of η^2 is symbolized by h^2.

To approximate the regression curve without assuming any relationship between the two variables, a piecewise linear regression is commonly used. In particular, Lopes da Silva et al. referred to a piecewise approximation of the regression curve by partitioning the independent variable, x_n, into L bins and connecting the points $(\overline{y}_{n,i}, x_{n,i}|_{mid})$ in the scatter plot of y_n as a function of x_n, where $\overline{y}_{n,i}$ is the average value of y_n in bin i and $x_{n,i}|_{mid}$ is the midpoint of x_n in bin i. Let the function $\hat{y}_n(x_n)$ be the estimation of y_n from x_n. The non linear correlation coefficient between x_n and y_n is calculated as:

$$h^2_{xy} = \frac{\sum_N (y_n - \overline{y}_n)^2 - \sum_N (y_n - \hat{y}_n(x_n))^2}{\sum_N (y_n - \overline{y}_n)^2} \tag{2.19}$$

The non linear correlation coefficient ranges between 0, when the two variables are independent, and 1 when one is completely determined by the other. Also, it is a symmetric measure if the relationship between the two signals is linear and asymmetric, i.e. $h^2_{xy} \neq h^2_{yx}$, otherwise. Also, one can calculate h^2_{xy} for different lags of y_n and conclude a possible causal relationship using the lag that maximizes the coefficient.

A summary of the methods mentioned here is shown in Table 2.1. Besides these ones, there are several other metrics of FC and EC, such as Granger causality [61] and phase locking value [14]. Though, one issue concerning any statistical method used to measure

Table 2.1: Methods used to calculate FC and EC.

Methods	**Model**	**Domain**	**Directionality**
Correlation	Linear	Time Domain	Undirected
Coherence	Linear	Frequency Domain	Undirected
Mutual Information	Model-Free	Time Domain	Undirected
Transfer Entropy	Model-Free	Time Domain	Directed
h^2	Model-Free	Time Domain	Directed

brain connectivity is the possibility of extrapolating from it a method for measuring its value between two neurological signals while controlling for the others, effectively avoiding spurious dependencies that come from a simultaneous dependency on other neurological signals, a characteristic only some possess.

Studies using linear methods still greatly outnumber the ones using non-linear methods[1] and a number of reasons can be associated with that. The difference between the number of available tools to measure each method, which is much higher for linear methods, might be one of the reasons. It might also be for comparison purposes or historical reasons. In addition, linear methods are less complex and have been associated with better robustness to noise (for a brief review on this matter see [20]).

From image acquisition to the determination of brain connectivity some steps might come into place to organise, spatially label or reduce the amount of information to be dealt with. The next Section will briefly cover this matter.

2.2.4 Image Registration and Brain Parcellation

Every brain is different. While some might be reasonably similar, others, like the ones from adults and children, are quite distinct. Due to this variability, defining regions in the brain is not as easy as applying the same three dimensional mask on every brain. Besides inter-subject variability, imaging techniques such has fMRI have low image resolution which hinders the identification of some brain constituents. Efforts to solve both these issues have been boosted in recent years by the fast increase in computational power and *in vivo* imaging techniques. Furthermore, not only is a consistent method needed to define regions in the brain, but the regions should also not be defined randomly. Image registration in a standard space and parcellation with brain atlases are two common preprocessing steps used to diminish anatomical variability and to consistently define meaningful brain regions.

[1] A search on PUBMED for papers using linear correlation as a connectivity measure was made with the terms: ("brain connectivity" OR "functional connectivity") AND (correlation OR cross-correlation) NOT ("non-linear correlation" OR "non linear correlation") revealed 325 results in 2016 while a search for papers using non-linear methods to measure connectivity with the terms: ("brain connectivity" OR "functional connectivity" OR "effective connectivity") AND ("Granger causality" OR "mutual information" OR "transfer entropy" OR "phase locking value" OR "generalized synchronization" OR "phase synchronization" OR "non-linear correlation" OR "non linear correlation") revealed 102 results in the same time period.

Image registration is made, for instance, using an algorithm that iteratively applies a transformation model to an anatomical or a functional image (the source or moving image) until a specific threshold of correspondence with a reference (or fixed) image is met [58]. Registration of brain images is usually made into a standard high resolution brain template or atlas to allow not only functional localisation but also comparison between subjects, by diminishing the inter-subject brain anatomy variability. Some brain atlases automatically label different regions of the brain, such as the widely known Automated Anatomical Labeling (AAL) atlas [141] which defines 116 regions. In this study, what interests us mostly is what the common practice in fMRI data is and as such, we shall now focus on image registration and brain parcellation of fMRI volumetric data. For general information on brain atlases, templates and their associated image and volume registration techniques, the reader is referred to e.g. [58, 143].

As previously stated, functional imaging, whether from fMRI, PET or Single-Photon Emission Computer Tomography (SPECT), deals at a fundamental level with the trade-off between temporal and spatial resolutions. Image registration relies on the quality of the source image for an accurate identification of the features needed to analyse its correspondence with the reference image. Because of the temporal resolution needed for functional imaging, and particularly for fMRI, the spatial resolution of these techniques does not serve this purpose in most cases. Nonetheless, functional to anatomical registration i.e. registration of a subject's functional image to his corresponding anatomical image, can still be accurately made because both images represent the same anatomical structure. Taking advantage of this, functional to template or atlas registration is possible by registering the functional images of interest to an anatomical image of the same brain and afterwards by applying the transformations needed to pass the high-resolution anatomical image to the wanted template or atlas, to those functional images [59]. Once registered to an atlas, or to an atlas space, the fMRI volumes become effectively parcelated or ready to be parcellated.

For FC or EC studies, the brain should ideally be parcellated in such a way that the time series of each voxel somewhat represents the general activation pattern [154] or the FC patterns [33] of the region in which it is included. Because every subject has its own activity and connectivity patterns, constructing an atlas based on this criterion is not an easy task, but some attempts have been made e.g. [37]. When using the statistical methods mentioned in Section 2.2.2, calculating the dependencies between the time series of every voxel of the brain in the fMRI volume can be computationally expensive. Thus, combining the time series of each region's voxels is a way of effectively diminishing the dimensionality of the data and, therefore, the computational cost of the process. Typically, the time series of each region is determined by averaging the time series of every voxel that constitutes it.

In summary, image registration normalizes each brain to a given template, allowing the automatic definition of a given number of Regions Of Interest (ROIs). The time

series of each Region Of Interest (ROI) can, then, be extracted by voxel-wise time series averaging and the subject's brain connectivity determined by calculating the statistical dependencies between each extracted ROI with a given multivariate analysis method. Thus, if a brain is constituted by n voxels in a given fMRI volume and the used atlas has m ROIs, after time series extraction for each ROI we pass from n, to m time series. The dependency values between all n time series ($n \times n$ values to be calculated) can be organized in a matrix, called FC matrix. Thus, by extracting a single time series for each ROI, the number of dependency values to be calculated is reduced by $n^2 - m^2$ values.

Even under the same conditions, each person has its own structural and connectivity patterns and one of the paradigms in neuroscience is to find a way of using this information to find consistent differences between groups of subjects, that would allow technicians to use these techniques for diagnostic and prognostic of neurological diseases. One possible technique that can be used to find group differences in the patterns of brain connectivity is ML.

2.3 Machine Learning

In many ways computers are like brains, both are systems that constantly receive input stimuli, convert them into electrical signals and process that information for storage and/or to produce an output. Despite their similarities, one area in which brains and computers are still quite different, is in the process of using the information they receive as experience to *learn* something new. *ML* can be described as the area of computer science responsible for making methods capable of detecting patterns in data and use them as experience to make future predictions [99, pp. 1-2].

Learning can be supervised, unsupervised, semisupervised or by reinforcement. In this text the first two types are going to be briefly described but the focus will fall mostly on the first. What distinguishes the type of learning is how the learning set, or the training set data is presented to the implemented algorithm. *For supervised learning*, the sample is given as a pair $\{\mathbf{x}_i, y_i\}$, $i = 1, ..., N$, $\mathbf{x}_i \in \mathbb{R}^n$ where N is the number of data points, usually subjects in neuroscience, n is the number of features per data point such as the age of each subject and y_i is a categorical or real-valued variable, called the label of the corresponding data point. The goal of supervised learning is to find an association rule between the points in the input feature space $\mathbb{R}^n$ and values of y_i. Thus, if a prediction is going to be made following those association rules, that prediction is of the same type as the label y_i. This means that the prediction can only be one of the values of y_i if y_i is categorical, or a real number if y_i is real-valued. The former case, in which y_i is categorical, is called a classification task and the latter is called a regression task. When the label is not provided and the task is to group data points that share a given number of features, the learning is said to be unsupervised.

2.3.1 Supervised Machine Learning Algorithms

Classification tasks generally have the same work-flow: data gathering, assembly of representative training, validation and test sets, choosing a learning algorithm for classification, parameter tuning in the validation set and finally model evaluation in the test set. Several algorithms can be used for supervised learning, some are instance based e.g. K-Nearest Neighbours (K-NN) [85], others are model based e.g. logistic regression [99, pp. 21-22]. In neuroscience, Support Vector Machine (SVM) [35] and Deep Learning (DL) algorithms are commonly used for classification and regression tasks (see for instance [90, 130]). Even though DL is of great interest for neuroscience, in this text it is not going to be covered because it was not used in this study. For a review of DL see e.g. [86] or for a book e.g. [60]. Next, SVMs are going to be introduced as well as ensembles of decision trees, which have been applied with success in classification tasks.

Support Vector Machines

In 1995, Cortes and Vapnik [35] introduced a new supervised and model based learning algorithm which draws linear decision boundaries based on the so called large margin principle. It was introduced for two class classification problems but has since been generalized to multiclass and regression problems, however, here it is only going to be considered the two class case. This algorithm was then called *support vector networks* by the authors and is now referred to as *SVM*. The previously mentioned large margin principle states that, of all decision boundaries that linearly separate two classes in the feature space, the best is the one that maximizes the margin between the boundary and the closest point to it, i.e., the one that maximizes the distance, r, to the closest point, measured orthogonally from the boundary defined by the discriminant function [99, p. 501].

In model based supervised learning, the discriminant function is a function, say $f(\mathbf{x}_i, \alpha)$, where α are its free parameters, that maps the values in the feature space to estimated label values: $\hat{y}_i = f(\mathbf{x}_i, \alpha)$. The ultimate goal is to find the values of α for which $f(\mathbf{x}_i, \alpha)$ maps every data point to the correct label, while providing a good generalization to new data points. For two class classification, the sign of the discriminant function decides on which side of the boundary each data point stands (positive value corresponds to a given label and a negative value to the other label; the decision boundary is defined by $f(\mathbf{x}_i, \alpha) = 0$). As previously said, in an SVM algorithm, the discriminant function is linear $f(\mathbf{x}) = \mathbf{w}^T \cdot \mathbf{x} + b$, where $\mathbf{w}$ is a vector normal to the hyperplane and b is the translation constant that controls the distance from the hyperplane to the origin. In this algorithm, the discriminant function should not only correctly map every point, but also, following the large margin principle[2], draw a decision boundary that maximizes r or,

[2]For the sake of clarity, the margin of the decision boundary is the set of points of the input feature space that satisfy the condition $|d| \leq r$, where d is the orthogonally measured distance from a given point to the decision boundary.

differently put, that minimizes $\|\mathbf{w}\|$ given that $r = f(\mathbf{x})/\|\mathbf{w}\|$. If we assume $y_i = 1 \vee y_i = -1$ and $y_s f(\mathbf{x}_s) = 1$, where s is the index of the data point closest to the decision boundary, we can mathematically express the initial objective of the SVM algorithm as [25]:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t.} \quad y_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1 \; \forall i \tag{2.20}$$

The minimization is on $\frac{1}{2}\|\mathbf{w}\|^2$ and not on $\|\mathbf{w}\|$ to simplify the derivative, which becomes just $\mathbf{w}$, and to make it differentiable at $\mathbf{w} = 0$ [57, pp. 145-165]. The data points that define the largest margin size are called *support vectors* and are the ones with index s, for which the equality in 2.20 holds.

There are two problems with an algorithm that only applies the large margin principle to a linear discriminant function, while having the correct value of $sign\{f(\mathbf{x}_i)\}$ for every $\mathbf{x}_i$. The first problem is that most classes are not linearly separable. In such cases, the constraint imposed on the sign of $f(\mathbf{x}_i)$ yields no solutions to the minimization problem (the constraint is never satisfied). To overcome this, Cortes and Vapnik introduced a slack variable, ξ_i, that measures how much $\mathbf{x}_i$ can violate the pre-defined margin. In the previous binary context, $\xi_i = 0$ if $\mathbf{x}_i$ is on the correct margin boundary or outside the margin but on the correct side of the decision boundary and $\xi_i = |y_i - f(\mathbf{x}_i)|$ otherwise [99, p. 501]. The objective is now to minimize $\frac{1}{2}\|\mathbf{w}\|^2$ and the margin violations, mathematically:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \quad \text{s.t.} \quad \xi_i \geq 0,\; y_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \tag{2.21}$$

where C is a constant.

The construction of the optimal hyperplanes is a convex quadratic programming problem [35] and the steps to its solution are not going to be covered here.

It was said that there were two problems with the mentioned algorithm, the first can be solved with the introduction of slack variables, but the discriminant function is still linear. To produce non-linear discriminant functions in the original feature space one might do a non-linear mapping of $\mathbf{x}_i$ from the original feature space to a high dimensional feature space and solve equation 2.21 in that high dimensional space. Computationally, however, this process can be very expensive. One way of going around this issue is by using the so called *kernel trick* in the corresponding Lagrangian dual problem which has the same solution of the primal problem (equation 2.21) in the given conditions[3]. The dual problem can be shown to be [25]:

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T \cdot \mathbf{x}_j - \sum_{i=1}^{N}\alpha_i \quad \text{s.t.} \quad \sum_{i=1}^{N}\alpha_i y_i = 0,\; 0 \leq \alpha_i \leq C,\; \forall i \tag{2.22}$$

where α_i and α_j are the Lagrange multipliers. After finding the minimal α one can calculate $\mathbf{w_{min}}$ and b^{min} using [18, pp. 334-335]:

[3] The inequality constraints and the objective function are convex and the former are continuously differentiable [57, pp. 145-165].

$$\mathbf{w_{min}} = \sum_{i=1}^{N} \alpha_i^{min} y_i \mathbf{x}_i \tag{2.23}$$

$$b^{min} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left(y_i - \sum_{j \in \mathcal{S}} \alpha_j y_j \mathbf{x}_i^T \cdot \mathbf{x}_j \right) \tag{2.24}$$

where $\mathcal{M}$ is the set of indices of the data points that satisfy $0 \leq \alpha_i \leq C$, $\mathcal{S}$ is the set of indices that correspond to the support vectors, which in this case are the ones that satisfy $y_i f(\mathbf{x}_i) = 1 - \xi_i$ and $|\cdot|$ is the cardinality of the set.

When using a mapping function, say $\phi(\mathbf{x})$, the dot products in equations 2.22 and 2.24 can be replaced by $\phi(\mathbf{x}_i)^T \cdot \phi(\mathbf{x}_j)$. The kernel trick consists in using a Mercer kernel function, $\kappa(\mathbf{x}_i, \mathbf{x}_j)$, to calculate the dot product $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \cdot \phi(\mathbf{x}_j)$ only using the original vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ [99, p. 481]. This means that even if the function $\phi(\mathbf{x})$ maps to a very high dimensional space, the computational cost is similar to the linear case in the original feature space.

The polynomial and the Radial Basis Function (RBF) kernels are two examples of Mercer kernels. The former allows one to work with mapping functions that originate features by multiplying the original features with each other and the latter allows one to work with functions that map to infinite dimensional spaces [99, pp. 481-482]. RBF kernels are widely used in SVM algorithms and consist of Gaussian kernels of the form:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = e^{\left(-\gamma_{SVM} \|\mathbf{x}_i - \mathbf{x}_j\|\right)^2} \tag{2.25}$$

where γ_{SVM} defines the width of the gaussian function.

It should be noted, lastly, that the determination of the discriminant function depends only on the support vectors and that its shape depends on the parameter C of function 2.21 and on the parameters of the kernel. High C values result in fewer margin violations but smaller margins and low C values result in the opposite. Because the decision boundary only depends on the support vectors this algorithm is memory efficient. Also, because of this, its performance might not be affected by the number of features [91].

Ensembles of Decision Trees

Classification based on a decision tree is made as follows[4]. The most discriminant feature of the original feature set is estimated and used as the root of the tree, from where the classification starts. By looking at the values of the root feature for each class, a self-complementary set of equality and/or inequality rules is put together with the goal of separating the existing classes as well as possible, using only the root feature. The process of creating rules from a given feature is called branching because each rule is going to correspond to given branch of the tree. Based on the separation of instances motivated by

[4] To simplify the explanation, the text refers to non oblique decision trees, which means that each node refers to a single feature. If interested in oblique decision trees, the reader is referred to [68].

the first branching, a class might be attributed to every data point that follows that rule, graphically, this corresponds to having a "leaf" at the end of the branch that corresponds to that rule. Instead of class attribution, a new set of rules might further divide one of the previously established fractions. This is made by creating a new set of rules based on the values of the feature that best discriminates the classes in the corresponding fraction. Again, graphically, this corresponds to the root branching to a node (the new estimated feature) and that node further branching into new nodes or leaves. This process can be repeated until every instance is correctly classified.

It is easy to identify the main issues concerning a decision tree. The first is node estimation, that is, the process of finding the most discriminant feature. There are several metrics used to estimate the most discriminant feature given a sample. Some of the most commonly used are Gini impurity [23] and information gain [113]. There are two other issues: how to derive the rules to separate the classes and the tendency of decision trees for overfitting. If one allows a decision tree to indefinitely grow, all instances of the training set, $\mathbf{x}_i$, will be correctly classified as long as the label of $\mathbf{x}_i$ is unique for all i. There is, however, a good chance of a fully grown decision tree to be overfitting the sample. A classification algorithm, or classifier, is said to overfit a given subset of a sample, if there is a different learned hypotheses that classifies that subset with a larger error but classifies the whole sample with a smaller error [91]. Some techniques, such as limiting the growth or pruning the fully grown tree [43] can be used to reduce the overfitting problem.

Ensemble methods combine the classification of several algorithms to make a final prediction. In case the combined algorithms are decision trees, these classification methods are called ensembles of decision trees. Examples of these type of classifiers are AdaBoost [50], Gradient Tree Boosting [52] and the Random Forest (RF) algorithm. As previously said, a decision tree cannot be grown to indefinite complexity. While pruning and growth limitation improve generalization to new datasets, they most often result in much lower accuracies in the training set. Ensembles of decision trees are an attempt at allowing high complexity, enabling, thus, good accuracy in the training set, while providing good generalization rules [31]. One of these methods, RFs, introduced in 1995 by Ho [70] and complemented by Breiman in 2001 [22], was used in this study.

In short, RFs combine the predictions of several decision trees, each called an estimator, constructed using randomly selected subspaces of the original feature space in each node [22]. This creates a group of decision trees, each with its own set of classification rules. Also, a technique called bagging is combined with a random feature subspace for further variability of the classification rules for each estimator. Bagging consists in drawing subsets from the training set with replacement [21]. Each estimator is then trained in a new training subset. The ensemble's prediction is usually made by voting of each estimator [70].

To make predictions, decision trees and RFs often use only a subset of the total number of features, also, they can use some features more than others or use some closer to the root

and others closer to the leaves, which means that they establish different importances for different features. This allows one to use decision tree based methods for feature selection, a topic to be developed next.

In neuroscience, it is common to have a large amount of information about a small number of subjects. In classification or regression tasks this translates to very high dimensional feature spaces and few instances. The performance of classifiers usually tends to degrade as data dimensionality increases [73], because the density of data points in the feature space decreases with increasing number of input dimensions. For instance, using the distance between data points as a similarity measure can possibly no longer be meaningful in high dimensional spaces because in such cases the ratio between the distance of an object to its nearest and its farthest neighbour approaches one [6]. Another consequence of high dimensional feature spaces is higher risk of sample overfitting [71]. The difference in the behaviour of data in high dimensional spaces in respect to its behaviour in low dimensional spaces is usually referred to as the curse of dimensionality [16].

2.3.2 Dimensionality Reduction and Feature Selection

Solutions to the curse of dimensionality in ML consist in dimensionality reduction and feature selection. Dimensionality reduction lies on projecting the data into a lower dimensional hyperplane while keeping the majority of the information. Probably the most common algorithm for dimensionality reduction is Principal Component Analysis (PCA) [77, 126], which finds a new basis to express data by making linear combinations of the original basis, with the assumption that important structures have high variance values. The components of the new basis are organized by descending order of importance, being the first, thus, the one that explains the largest ammount of variance. Dimensionality reduction techniques can reduce redundancy in data and possibly noise, however, most of the times, the new representation makes it difficult to draw conclusions about the influence on classification of a given feature of the original feature set. The goal of feature selection is to find (possibly the smallest) subset of features that most benefits classification.

Supervised feature selection techniques (feature selection techniques that use label information) can be divided into three types [101, 119]: filter techniques, wrapper techniques and embedded techniques. Filter techniques are based on statistical measures such as t-tests and ANOVA, that test the null hypothesis of the mean values of each class of a given feature being equal. Both of these statistical tests are very popular and have been used for feature selection in neuroscience. The major downside of filter techniques is that they consist mostly in univariate methods. It is expected for most of the patterns present in high-dimensional data to arise from the relationship between 2 or more variables. Thus, it is possible for one feature to not give any insight on how the classes differ from

each other, by itself, but that its relationship with other variables allows us to perfectly separate the classes.

Wrapper techniques for feature selection are based on classifiers and can be of forward selection or backward elimination [101]. These include iterative methods based, respectively, on gradual incrementation of features into the feature space in each iteration or gradual elimination of features of the original feature space, in each iteration. The choice of which features to add or remove in each iteration is usually made by comparing the classification performance using an initial feature space (possibly empty in forward selection), with the classification performance using the initial space plus or minus a given subset of features. Rules to stop the iterative search can be based on a previously fixed goal for the number of features or the cycle can continue while a given threshold of classification accuracy increment is being surpassed in each iteration.

Lastly, embedded feature selection techniques use models of supervised ML algorithms or make restrictions on them. One example is to force, in a given classifier model, the weights associated with certain features to be zero (L1 regularization). Embedded feature selection techniques include the very popular Least Absolute Shrinkage and Selection Operator (LASSO) technique [137, 138], the Elastic Net [159] and the partial least squares method.

2.3.3 Classifier Evaluation

Following classification, one would want to estimate how well the classifier would perform outside the data used to derive the decision boundary, that is, an estimation of the classifier's *generalization ability*. To achieve this, k-fold cross-validation is usually used due to its simplicity and because it guarantees that every data point is used once to test the classifier. In short, k-fold cross-validation consists in randomly dividing a sample into k fractions and making k classifications using in each one a different fraction as the test set and the other k−1 fractions as the training set. The training set is the fraction of the sample used to derive the decision boundary and the test set is the fraction of the sample used to evaluate the generalization ability of the classifier. Typical values of k are 5 or 10. When k = N, where N is the number of data points of the sample, the training set is the whole sample minus one element which serves as the test set. The particular case k = N is called Leave-One-Out (LOO) cross-validation. Because the number of elements in the training set is as high as it can possibly be, LOO cross-validation is expected to result in better predictions. Also, one advantage of LOO cross-validation over any other kind of k-fold cross-validation is the elimination of performance variability across trials coming from the random splitting of the data that happens in the cases where k < N. Thus, comparison of classification performances between studies or between different classifiers on the same sample are more reliable when LOO cross-validation was used in both cases for model evaluation.

To get a general idea of the classifier's performance in the test set, the simplest and

most intuitive measure one can use is the prediction accuracy, which is the fraction of correct predictions made by the classifier. Accuracy is almost always reported in classification studies and is often the only measure reported in abstracts. Probably, the major downside of accuracy is its bias regarding the sample skewness. A classic example that illustrates this is a two class sample with 90 elements of class A and 10 elements of class B. A classifier that predicts always class A would achieve an accuracy of 0.9. Thus, when analysing accuracy reports one should always take in consideration how the classes are balanced.

In two class classifications, a more complete view of the classifier's performance can be achieved using a confusion matrix, which reports the number of True Positives (TP) and False Positives (FP) and the number of True Negatives (TN) and False Negatives (FN) (type I and II errors)[5]. In fact, it is such an important element in performance analysis that most measures, including accuracy, are simply ways to summarize or highlight some of the information it contains.

Precision, also called positive predictive value, measures how many data points predicted as belonging to the positive class actually belong to the positive class:

$$Precision = \frac{TP}{TP + FP} \tag{2.26}$$

The True Positive Rate (TPR), also called sensitivity or recall, measures how many data points belonging to the positive class were correctly predicted:

$$TPR = \frac{TP}{TP + FN} \tag{2.27}$$

In medical context, if the negative class is considered as the healthy class, good TPR is very important since the goal is to completely eliminate FN. Both of these measures are defined to evaluate estimations on the positive class, but similar measures can be defined for the negative class. The *f-measure* is used to integrate information about the classifier's precision and TPR and is defined as their harmonic mean [98, p. 283]:

$$Fm = 2\frac{precision \cdot TPR}{precision + TPR} \tag{2.28}$$

Again, the f-measure is biased because it only takes into account variations in the positive or the negative class, but not in both simultaneously. One way to achieve a measure that can evaluate predictions on both classes and that takes into account the data's skewness is to average f-measure and its correspondent when the positive and negative classes are interchanged. This average can be weighted or not, in which cases this measure is called micro-averaged f-measure or macro-averaged f-measure [140], respectively. In two class classification, micro-averaged f-measure equals accuracy. If a random prediction is going to be made, while taking into account the weights of each class, then, the expected

[5] In two class classifications, one class is usually called the positive class and the other the negative class, hence the true/false positives/negatives notation

value of the macro-f-measure is 0.5. There are, however, two setbacks of using macro- or micro-f-measures to evaluate the classifier's performance. First, they are not easy to interpret and second the macro-averaged f-measure is not defined when the classifier always predicts the same class.

Several other measures can be defined using the confusion matrix and most can be found in [110]. Also, by changing the discriminant function threshold, TPR vs False Positive Rate (FPR)[6] curves, also called Receiver Operating Characteristics (ROC) curves [161], as well as precision vs TPR curves, are very common to evaluate binary classifications. The Area Under the Curve (AUC) is used to summarize the information about the ROC curve and, as the name indicates, measures the area under the ROC curve. A perfect classifier has an AUC = 1.

One aspect that should be taken into account when evaluating a classifier's performance is that when an extrapolation to "real world" applications is intended, often not all prediction errors have the same importance. An example is in classifications of patients vs Healthy Controls (HC), as was previously mentioned. A cost can be attributed to each error and organized in cost matrices, and, then, the cost of a classifier's prediction can be determined using that information.

The generalization ability and performance of a classifier depend on the value of its parameters. Parameters should be chosen to optimize generalization to new data, not the performance on the training data, though both are intimately connected. Parameters are specific to each supervised algorithm but usually at least one per algorithm controls the complexity of the model. Even in linear models one can control the model complexity by forcing the decision boundary to pass through the origin of the feature space or its slope to be limited within a range of values. In an ideal case, a perfect separation in the training set would mean a perfect generalization rule. Though, that is almost never the case due to the statistical outliers. If the training set is not representative of the whole sample, the decision boundary built based on it will not generalize well to new data. The optimal parameters are the ones that allow the model to ignore the outliers and focus on the data points that represent the whole sample.

Choosing the optimal parameters is often a heuristic task. First, one should choose which performance measure to optimize. This can vary depending on the application but most commonly the chosen one is accuracy. If a trial and error approach is chosen to tune the parameters, one has to decide in which fraction of the whole sample to do this. The vast majority of the time this is done in the test set. It is easy to understand why this leads to a positive bias in the evaluation of the classifier. By using the test set to find the optimal parameters, one is using information of the test to draw a decision boundary. When the classifier is tested in new data, it is expected to have a lower performance than that estimated in the test set because the parameters are chosen to optimize results in the test set [110]. In this context, one can conclude that k-fold cross-validation leads to biased

[6]The FPR is defined in the same way as the TPR but with the number of FP in the numerator

results. To avoid this, a different fraction of the whole sample should be used to test the data and to tune the parameters. This new set is usually called the validation set. One alternative to cross validation is to randomly split the data into a test set and a different subset. In this subset one could fraction it in k parts and do k-fold cross-validation to tune the parameters in the now called validation fold. The training would be done in the other k−1 folds. The results achieved with this technique rely too much on the first split which can be composed of data points harder or easier to classify than average. There is, however, a computationally costly alternative to this, called nested cross-validation [98, pp. 272-275].

Nested cross-validation randomly divides the data in K folds, then, at a time, each of these K folds serves as the test set. The other K−1 folds are grouped and divided in k folds and regular k-fold cross-validation is used for parameter tuning. After classification in the test set with the parameters found in cross validation, a different fold of the set of K is chosen as the test set and the parameters are again tuned by k-fold cross-validation in the other K-1 folds. The procedure is repeated for all K folds. K-by-k nested cross validation is, thus, a series of k-fold cross-validations inside a K-fold cross-validation. Besides its computational cost, this technique provides K different classifiers and the choice of which of them to use in new data is not straightforward. Besides that, in small samples, the training and test set can become too small to provide reliable classifiers and performance evaluations, respectively.

Finally, facing the variability of the classifier's performance one could use statistical tests to measure the deviation of its distribution to a random one or to compare it to other classifications [134].

Chapter 3

Literature Review

Unlike most medicine branches, psychiatry still classifies mental disorders based on external clinical signs [78]. The main reason for this comes from the difficulty of finding cause-effect relationships in such disorders [39]. Functional and structural connectivity have been at the center of a significant portion of the efforts to gather the knowledge needed to close this gap, with the hypotheses that some of these neurological and psychiatric disorders are caused by disruption of connectivity between brain regions [54, 69]. As already mentioned, one possible path to find connectivity pattern differences between groups is classification through ML.

FC and EC, in particular, have been used extensively as feature extraction techniques from neuroimaging data for classification purposes. Using features originated from FC and EC, several studies using ML techniques to separate HC from patients, have been done for many mental disorders: autism [8, 102, 109], schizophrenia [9, 27, 32], Major Depressive Disorder (MDD) [76, 155, 156], Alzheimer's Disease (AD) [75, 157], Obsessive-compulsive Disorder (OCD) [12, 123, 125], dyslexia [51] and ADHD [41, 56, 88, 115, 148], though, in almost all of them, the constructed classifiers were evaluated in small test sets.

The first classification of ADHD vs controls using rs-fMRI was made by Zhu et al. [158] using regional homogeneity, which is a local measure of synchrony of brain activity, in 9 ADHD patients and 11 HC, achieving 85% accuracy. Since then, other studies using rs-fMRI have been made, some already referred previously [41, 56, 115, 148], as well as studies using structural MRI [74, 107] and task fMRI [66, 67].

3.1 Classification of subjects in the ADHD-200 and ACPI databases

To allow studies in larger datasets and testing without bias, a global competition [4, 15] called ADHD-200 occurred in 2011 which made available preprocessed rs-fMRI and structural MRI data of 973 individuals with a predefined test set [15]. Data from this dataset was made available by several acquisition sites, particularly by: the KKI (train and test sets), the NeuroIMAGE (train and test sets), the NYU (train and test sets), the OHSU (train and test sets), the Pittsburgh (train and test sets), the Peking (1- train and test sets 2- and 3- train set), the BROWN (test set) and the WashU (train set) sites. Accuracy performances resulting from the competition were around 61% [44, 153].

Later classifications using ADHD-200 rs-fMRI data almost always used only a subset of the whole dataset. Some examples of authors that classified ADHD vs HC using data from the ADHD-200 dataset include: Ge et al. which achieved a maximum of 90% accuracy on a balanced dataset with 40 subjects sampled from the NYU dataset, using LOO cross-validation [56]. These authors used FC calculated using Pearson's correlation coefficient to extract the features and used an SVM classifier with a RBF kernel. Qureshi et al. [115] used voxel-wise FC for feature extraction and achieved 71% 3-way (HC vs ADHD-C vs ADHD-I) accuracy with a 10-by-10 nested cross validation using a hierarchical extreme machine learning classifier [135] and 67% using an SVM classifier in a sample with 60 ADHD patients and 30 HC. Judging by the results available in the literature, it seems that using the ADHD-200 test set to evaluate the classifier has a big impact on its performance measures. For instance, Hao et al. used the training and test sets of three acquisition sites and classified them independently [64]. The used sample had ADHD+HC balances in the test set of 29+12, 23+27 and 3+8 subjects, and 4-way (HC vs ADHD-C vs ADHD-I vs ADHD-H) classification accuracies in each were 49%, 54% and 73% in the datasets from the NYU, Peking-1 and KKI sites, respectively. Dey et al. [42] achieved 4-way accuracy values of 54% in the KKI's test set and 48%, 82% and 59% in the NeuroIMAGE, OHSU and Peking-1 test sets from the ADHD-200 sample. Table 3.1 presents a summary of the classification studies that used data from the ADHD-200 database.

Another publicly available dataset with rs-fMRI data from subjects with ADHD and HC is the ACPI dataset [1]. Studies using this dataset are scarcer than the ones using data from the ADHD-200 sample and, in fact, only one study [93] reported ADHD vs HC classification results. This study used 126 subjects, 86 with ADHD, from the Multimodal Treatment of Attention Deficit Hyperactivity Disorder 1 (MTA 1) subsample [95]. These authors achieved a macro-averaged f-measure of 0.51 using FC calculated with the Pearson's correlation coefficient and 0.58 using FC calculated with dynamic time warping [93], while using an SVM classifier and achieved 0.44, and 0.60 when using a LASSO classifier.

From the previously mentioned results, one can conclude that some differences in brain activity between patients with ADHD and healthy subjects have been found, in

particular when analysing brain connectivity derived from rs-fMRI. Some differences include the connectivity between frontal areas and the cerebellum [41] and also regional homogeneity in the prefrontal cortex and the anterior cingulate cortex [158]. Nonetheless, good generalization ability to larger and more heterogeneous samples has not been frequently achieved, as can be concluded from results using the predefined ADHD-200 test set. Up to now, the exception seems to be a study by Deshpande et al. [41]. The authors used 744 rs-fMRI volumes from HC and 433 from subjects with ADHD, preprocessed with the Athena pipeline [11]. Then, they calculated connectivity between regions obtained with the atlas developed by Craddock et al. [37] using four statistical methods: Granger causality [61] (model order 5), kernel Granger Causality [92], correlation between probabilities of recurrences [116] and correlation-purged Granger causality [40]. A t-test was used to select the 200 most significant features of the four connectivity matrices. These features were used for classification with a Fully Connected Cascade (FCC) Artificial Neural Network (ANN) classifier. Using LOO cross-validation the authors reached 90% accuracy when classifying ADHD-C vs HC and when classifying ADHD-I vs HC. The authors also performed classification using an SVM classifier with an RBF kernel with "infinite" C and $\gamma_{SVM} = 11$. Unfortunately, results with the SVM classifier were only reported in a form of graph and, therefore, exact accuracy values are not available.

Results as the ones from Deshpande et al. lead one to think that making feature selection from a bigger set of features can have a significant impact on classification performances and also, if those features are biologically inspired, on the discovery of what biological causes underlie a given mental disorder. As seen in the theoretical background (Chapter 2), statistical methods to measure FC and EC focus on or highlight different aspects of the statistical dependency between the activity of different brain regions, and combining them might have a positive impact on classification performance, as was the case in Deshpande's study.

Table 3.1: Summary of classifications using rs-fMRI in the ADHD-200 dataset.

	Test size ADHD+HC	**Features**	**FC**	**FS**	**Classifier**	**Evaluation**	**Accuracy*** **(%)**
ADHD-200 competition [44]	350+554	multi.	multi.	multi.	multi.	PTS	HCvsADHD: 61 (max)
Ge et al. 2015 [56]	20+20	ROI	PCC	yes	SVM	LOO-CV	HCvsADHD: 90
Qureshi et al. 2017 [115]	(30+30)+30	ROI	PCC	yes	SVM, H-ELM	10-by-10-CV	HCvsADHD-IvsADHD-C: 71
Hao et al. 2015 [64]	NYU, Peking-1, KKI	ROI	no	yes	DBaN	PTS	HCvsADHD-CvsADHD-IvsADHD-H: 49-54-73
Kuang et al. 2014 [81]	NYU, NeuroImage, OHSU, Pittsburgh	ROI	no	yes	DBaN	PTS	HCvsADHD-CvsADHD-IvsADHD-H: 37-44-81-56
Dey et al. 2014 [42]	KKI, NeuroImage, OHSU, Peking-1	GD	no	yes	SVM	PTS	HCvsADHD: 55-48-82-59
Deshpande et al. 2015 [41]	433+744	ROI	GC+CPGC+KGC+CPR	yes	FCC-ANN	LOO-CV	ADHD-C vs HC: 90, ADHD-I vs HC: 90
Wang et al. 2013 [149]	23+23	ROI	ReHo	yes	SVM	LOO-CV	HCvsADHD: 80

FS, feature selection; multi, multiple; PCC, Pearson correlation coefficient; GC, Granger Causality; CPGC, correlation-purged Granger causality; KGC, kernel Granger causality; CPR, correlation between probabilities of recurrences; GD, graph distances; ReHo, regional homogeneity; H-ELM, hierarchical extreme learning machine; DBaN, deep Bayesian network; FCC-ANN, fully connected cascade artificial neural network; PTS, predefined test set of the corresponding acquisition sites; ADHD-I, ADHD inattentive type; ADHD-H, ADHD hyperactive type; ADHD-C, ADHD combined type. *Values are in the same order the test sets were presented.

3.2 Comparison of Statistical Methods for Brain Connectivity Estimation

Comparisons on brain connectivity measured with different statistical methods are scarce and to the author's knowledge, in fMRI, it was only done systematically in simulated data. Smith et al. [127] provided an interesting discussion on how one can try to compare FC and EC measured by different methods considering the subjective nature of their definition. As was very well questioned by the authors: "If one method identifies a pattern of connectivity in a dataset that another method fails to find does this indicate the first method is superior? If two methods identify a similar pattern of connectivity in the same dataset, does this indicate the connectivity is more likely to exist?" [127, p. 2]. A different group [129] simulated several rs-fMRI data using Dynamic Causal Modelling (DCM) [53] varying the amount and type of noise in each simulation. To recover the simulated connections, 20 statistical methods were used including correlation and partial correlation, partial correlation via inverse covariance, coherence, general synchronization, mutual information, Granger causality, Patel's conditional dependence [105] and several Bayes net models. The authors concluded that in data without added noise, Bayes net models and both partial correlation methods perform better based on the results achieved in the simulated data, while Granger causality performed the worst. Patel's conditional dependence was the method that best recovered causality direction. In data with more that 50 nodes (would correspond to 50 voxels or ROIs in real data, for example) Granger causality results were not reported due to the implied computational cost, which seems to be one of the major downsides of this method. In terms of other variants, connectivity recovery was best for longer simulated time series and the most confounding factor was mixing of several ROIs' signals, which was an attempt to simulate the effect of defining ROIs that do not represent the actual functional boundaries. Also, the authors noted that measuring causality using lag-based methods should be done with caution since this kind of relationships can be blurred or created by specific haemodynamic patterns.

Similarly, Wang et al. compared several statistical methods in simulated fMRI data [147], again using DCM, and EEG data using a convolution-based neural mass model [94]. The authors used 42 statistical methods, most being directed and/or partial variants of correlation, coherence, Granger causality, mutual information, transfer entropy and non-linear correlation [89] (h^2). Using the correctly identified connections as true positives and the incorrectly identified connections as false positives, the authors plotted the ROC curve for each method and measured the AUC to evaluate performance. Each method was tested for different temporal window sizes, overlap between windows and other method-specific parameters in order to find their optimal values. Contrarily to the results of Smith's study, without noise, Granger causality based methods performed the best on fMRI data and also, Fourier based coherence and transfer entropy achieved good results even for small window sizes. Again, Granger causality was noted for its very high

computational cost when compared to every other method[1]. Linear relationships were also simulated and mutual information and h^2, two non-linear methods, did not perform well in those cases. Though, biological relationships seem to rarely be linear. Also, optimal size window when calculating FC or EC from fMRI data was higher than 100 s in all reported methods, and optimal fMRI session length was higher than 1000 s. No information about optimal model order for Granger causality or Multivariate Autoregressive (MVAR) based methods nor about optimal maximum delay for lag-based methods was given for fMRI data. Wang et al. attributed the difference between the results obtained with Granger causality in both studies, to the weak connection strength of the simulated connections in Smith's study or to the different parametrization used by them. In both studies, the most commonly used FC measure in neuroscience - correlation - performed reasonably well even in signals with simulated noise.

3.3 Impact of Feature Selection in Neuroscience

As was said in Chapter 2, feature selection techniques are expected to lead to better generalisation abilities if a good selection is made, because high dimensional data is not handled well by most classifiers and neither are "noisy" features, which tend to mislead them. This issue has been recently tackled by Chu et al. [29] for anatomical MRI studies, which concluded that in such cases feature selection does not always improve classification performance. fMRI data has even higher dimensionality and the need for dimensionality reduction in such cases might give more importance to feature selection techniques. Two examples that seem to confirm this hypothesis are the study from Craddock et al. [36] which introduced two feature selection techniques and verified that those methods diminished prediction error relative to no feature selection. The same was found by Wang et al. [150], which also introduced a new method for feature selection based on mutual information. Better evidence for this conclusion and particularly for classification in the ADHD-200 dataset was provided in [55]. The authors discussed the importance of feature reduction techniques in rs-fMRI data for classifications distinguishing ADHD vs HC groups and classified independently data from the KKI, NeuroIMAGE, NYU and Peking sites of the ADHD-200 training set sample, using an SVM, a K-NN, a naive Bayes, a perceptron and a C4.5 [114] classifier. Features used in classification came from FC matrices calculated using Pearson's correlation coefficient. All four feature selection techniques resulted in better overall classification performances measured using[2]: $(TPR + TNR)/2$, than what was achieved when no-feature selection was made. Particularly, maximum results were 0.60 when using a naive Bayes classifier plus LASSO feature selection for data acquired in the NYU site vs 0.49 using the same classifier but no feature selection,

[1] The authors estimated a polynomial approximation of order 7 for Granger causality and of order 2 for the rest when calculating the computational cost as a function of the number of simulated nodes.

[2] The True Negative Rate (TNR) is calculated similarly to the TPR but for the negative class.

0.87 in the NeuroIMAGE dataset using a C4.5 classifier with a Max Relevance Max Distance (MRMD) [160] feature selection vs 0.78 with a C4.5 classifier in the same dataset but without feature selection, in the KKI dataset, the maximum, 0.68, was also achieved using a C4.5 classifier and a Wilcoxon feature selection [119] vs 0.63 without feature selection. Results in the Peking dataset were not reported but the authors did mention that results were, contrarily to the previous cases, better without feature selection. In DL techniques, Vieira et al. [146] suggested that using feature selection before using a DL model seems counter intuitive due to the ability of DL models to select the best features out of raw data. Based on theory and the previously mentioned results, feature selection in rs-fMRI data and particularly in FC measures, generally seems to have a beneficial impact on classification, though, results might heavily rely on how well the selection was made.

Chapter 4

Materials and Methods

To make a comparison between classifying subjects using one or using several connectivity matrices, rs-fMRI datasets from two distinct databases were used, more specifically one from the ACPI [1] database and another from the ADHD-200 database [4, 15], both part of the 1000 Functional Connectomes Project (FCP) and the International Neuroimaging Data-Sharing Initiative (INDI) [28].

4.1 ACPI database

The scan parameters of the ACPI database's rs-fMRI sessions can be found on the project's website [2]. Preprocessed data from the MTA 1 dataset [95] of the ACPI database was used and included 125 subjects (all but the subjects with the following IDs: 28040, 28050, 28106 and 28119[1]), 101 males and 24 females, with ages between 21-27 years, 85 diagnosed with ADHD (68%). Preprocessing of the raw 4D rs-fMRI data had been made using a Configurable Pipeline for the Analysis of Connectomes (C-PAC) [26] and the Advanced Normalization Tools (ANTs) pipeline and consisted in the removal of the first five fMRI volumes, anatomical registration, tissue segmentation, functional registration in the Montreal Neurological Institute (MNI) space, functional masking, temporal bandpass filtering (0.01 – 0.1 Hz), motion correction, spatial smoothing and various nuisance corrections, more details can be found on the website [136]. Some variations of the preprocessing pipeline were available, namely: motion correction with and without scrubbing, combined with nuisance correction with and without global signal regression. Scrubbing would occasionally result in data with less time points than ROIs, a condition not supported by some statistical methods used to calculate FC matrices. Also, some evidence

[1] These subjects were not included because their data was not available.

was found that global signal regression introduces spurious anti-correlation patterns between regions [100]. Having this into account, the preprocessed data without scrubbing and without global signal regression was chosen. To build connectivity matrices with a reasonable number of weights, parcellated data was used. Of the parcellations available on the ACPI's website, AAL [141] was chosen, hoping that its high number of ROIs would yield a more accurate representation of the brain's FC.

4.2 ADHD-200 database

Unlike the data from the MTA 1 dataset acquired with a single scanning protocol, the ADHD-200 database consisted of raw rs-fMRI data from eight independent imaging sites each with its own scanning procedures with TR ranging from 1.96 – 3 s and inconsistent resting state instructions - for instance, in some sites the subjects were instructed to keep their eyes fixed in a crosshair, while in others they were instructed to keep their eyes closed. Information about the quality of the rs-fMRI time series was given (usable vs questionable) and all questionable subjects were discarded. Moreover, some subjects had made more than one session and, if more than one of them was classified as usable, then the first one made by the subject was kept and the rest discarded to remove intra-subject variability. The resulting dataset consisted of 661 rs-fMRI sessions from 661 subjects, with ages between 7-21 years, 234 diagnosed with ADHD ($\approx 35\%$). As was said in Chapter 3, in 2011 the ADHD-200 database was subject of a classification competition aiming at the discovery of novel biomarkers for ADHD [4, 15]. This was one of the main reasons for the choice of this database out of all the ones considered. Being subject of a classification meant that several results were already available, enabling later comparisons. For this competition, a predefined test set was established, containing 197 rs-fMRI sessions each from a different subject. Unfortunately, even though all of the included sessions were classified as usable, information about the class of some subjects was lacking and, therefore, for this study, those had to be discarded. The remaining 146 subjects, 66 diagnosed with ADHD ($\approx 45\%$), were, then, used as the test set for the classifier's evaluation and the aforementioned 661 subjects as the training set. The rs-fMRI sessions of the test set were provided by seven of the eight sites that provided data for the training set plus a different one, the age range in both the test and the training set was similar. Preprocessed data from these sessions was available through the Preprocessed Connectomes Project (PCP) [15, 111]. The downloaded data was preprocessed by R. C. Craddock with the most commonly used [15] Athena pipeline, based on tools from the AFNI [111] and FSL [7] software packages. Preprocessing of the rs-fMRI data included removal of the first four volumes, anatomical registration, functional registration in the MNI space, functional masking, motion correction and spatial smoothing. More details can be found on the website [11]. Two preprocessed versions of the data were used for this study: with and without temporal bandpass filtering (0.009 – 0.08 Hz), with the purpose of analysing the effect of filtering in classification. This filter is usually applied because BOLD signal FC

is believed to be reflected in such frequency ranges [19, 34]. As with the selected ACPI dataset, AAL parcellated data was used.

The choice of both of the previously mentioned databases was based on the availability of the preprocessed data and on their few but interesting differences. On the one hand, both include ADHD and control subjects, both were parcellated with the same atlas and also, in both the preprocessing steps were somewhat analogue. On the other hand, though, the number of subjects in the dataset selected from the ACPI database was much smaller than in the one selected from the ADHD-200 database (125 to 661+146 respectively) and, because of that, in the dataset selected from the ACPI database, some experimentation without too much time consumption was allowed. Also, data from the ACPI database was acquired with a consistent protocol and can, thus, be considered more homogeneous than the one from the ADHD-200 database. Finally, another reason for choosing two different databases with the same classes of subjects was the confirmation of any hypothetical findings.

4.3 Connectivity Matrices

FC matrices were built from the parcellated data using a batch version of the Multiple Connectivity Analysis (MULAN) open source toolbox [97], developed in MATLAB by Wang et al. [147], written for the purpose of this study. This software allowed for the calculation of 42 statistical methods between discrete, one dimensional signals. We shall follow here the division of these 42 methods into 7 families as presented in Wang's paper. The 7 families are: correlation, h^2, mutual information, transfer entropy, coherence, Granger causality and $\overline{\mathcal{A}}\mathcal{H}$. The h^2 family is composed of methods using the non-linear correlation coeficient and the $\overline{\mathcal{A}}\mathcal{H}$ family composed of methods using the frequency domain of MVAR models. For each subject, the Fourier and wavelet based coherence and the directed and non-directed versions of the statistical methods of each family but the last three mentioned before, were calculated between the time courses of all 116 AAL ROIs, including between a given region and itself. This resulted in 15 FC matrices, calculated from 11 statistical methods (three for each of the two methods of the coherence family and one for each of the other nine methods, see table 4.1). Each matrix had $116 \times 116 = 13456$ weights each. The mathematical description of the methods is present in the supplemental material of the previously referenced paper [147].

For a given statistical method, the yielded weights depend on the parameters chosen for its calculation. The correlation family, for example, includes the undirected bivariate Pearson correlation method which can be calculated for a window with a number of time-points ranging from one to the whole time series, thus, making the number of time-points per window one of the parameters of this method. For the sake of brevity, the reader is referenced to Wang's paper [147] and the MULAN toolbox [97] for the description of all of the parameters needed for the calculation of the aforementioned methods.

In table 4.1 the values used for each parameter are summarized. It should be noted that the FC matrices are heavily dependent on the chosen parameters as shown, for simulated fMRI signals, in Wang's paper. Taking into account that it would be infeasible to repeat the whole classification process for parameter optimization, a qualitative analysis of Wang's study was made to choose the parameters to use. Because the results in said paper generally improved with window size, small analysis windows were avoided and the whole time series was always considered (ACPI max: 179 time-points or 388.43 s; ADHD-200 max: 257 time-points or 503.72 s). There are also two added benefits of using a window that includes the entire time series, the first one being time consumption. Because we are dealing with 11 different statistical methods, there will be, at least, 11 FC matrices. Let us suppose, then, that combinations of 2 are going to be made from those 11 matrices. In that case, one would need to classify, at least, ${}^{11}C_2 = 55$ samples. However, if the analysis window is going to be considered as half of the whole time series, then, one would have at least twice as many, or 22, matrices and ${}^{22}C_2 = 231$ samples to classify, which would represent a 420% increase in time consumption. The second benefit of using the whole time series is the elimination of the overlap variable. Once there is more than one analysis window, how much one overlaps with another influences the calculated weights and we would wish to avoid dealing with choosing an overlapping value. Thus, taking into account the whole time series in a single analysis window seemed to be the best choice, even though the number of time-points considered often varied from subject to subject, which is not an optimal condition.

With the chosen parameters, the weights of the three matrices calculated with the bivariate wavelet based coherence method were equal and as such, only one of them was used in classification.

Two families were left out of the classification study: the Granger causality and the $\overline{\mathcal{AH}}$ families. In the former, the calculation of weights for model orders greater than two were very demanding computationally and quite time consuming. Also, the initial results for models of order two were fairly worse than the ones achieved with the other methods used. Thus, even though this family was the one that performed best in the study made by Wang et al., it had to be discarded for the final classification. Regarding the $\overline{\mathcal{AH}}$ family, the large number of methods that constituted it, and their heterogeneity[2] were the decisive factors for it being left out of the study.

4.4 Classification

At this point it should be remembered that the main goal of the present study is to make a comparison between the classification of subjects using a single FC matrix and the classification of subjects using a combination of FC matrices in the selected ACPI and ADHD-200 datasets. Besides that, it should also be pointed out that a comparison between

[2]Heterogeneity in the sense that its methods were not simple variations of each other.

Table 4.1: Parameter values for every method available in MULAN following the terminology in [97, 147].

Methods	**Parameters**		
BCorrU	MaxDelay = 12	—	—
BCorrD	MaxDelay = 12	—	—
BCohF	minfreq = 0.01 Hz	maxfreq = 0.1 Hz	stepfreq = 0.033 Hz
BCohW	minfreq = 0.01 Hz	maxfreq = 0.1 Hz	stepfreq = 0.033 Hz
BH2U	MaxDelay = 12	bins = 16	—
BH2D	MaxDelay = 12	bins = 16	—
BMITU	MaxDelay = 12	bins = 16	—
BMITD1	MaxDelay = 12	bins = 16	—
BMITD2	MaxDelay = 12	bins = 16	—
BTEU	MaxDelay = 12	—	—
BTED	MaxDelay = 12	—	—
			For all: wins = WTS

BCorrU, undirected bivariate Pearson correlation; BCorrD, directed bivariate Pearson correlation; BCohF, bivariate Fourier based coherence; BCohW, bivariate wavelet based coherence; BH2U, undirected bivariate h^2; BH2D, directed bivariate h^2; BMITU, undirected time domain bivariate mutual information; BMITD1, directed time domain bivariate mutual information by comparing the individual histograms to the joint histograms; BMITD2, directed time domain bivariate mutual information similar to BMITD1 but reducing the discretization bias; BTEU, undirected bivariate transfer entropy; BTED, directed bivariate transfer entropy; WTS, whole time series.

classifying with a single FC matrix calculated with different statistical methods is going to arise as a consequence of the process to achieve the main goal and that we will also be able to explore the impact of band-pass filtering in the frequencies expected to carry most of the BOLD signal's information, using both filtered and non-filtered data from the ADHD-200 dataset mentioned in Section 4.2. Finally, classification with and without prior feature selection shall also be compared.

The input features for classification are selected weights of a given matrix. Thus, for each matrix, there will be at maximum as many features as there are weights. First, a description of the classification process in the ACPI dataset shall be made and then one for both ADHD-200 datasets. To the trained classifiers was presented data with and without prior feature selection for all three datasets.

Classification was made using Python programming language and the scikit-learn package [106]. This package is the result of an open source project aiming at developing state-of-the-art ML algorithms [98]. The documentation for each algorithm is available on the project's website [122].

4.4.1 Classification in the ACPI dataset

To classify subjects as HC or as patients with ADHD in the ACPI dataset a model evaluation technique had to be chosen. Two options were considered, stratified k-fold/LOO cross-validation or k-fold nested cross-validation. Even though nested cross-validation avoids leakage of information to the test set, making it the best option in most cases, the number of subjects in the training, validation and test set were too small for the number of features used in classification. As such, regular 5-fold and LOO cross-validation were performed in the ACPI dataset. Data was classified using an RF and an SVM classifier, as implemented in the scikit-learn package, the latter with a Gaussian RBF kernel. For the RF classifier, a maximum number of 116 features per node was used, using Gini impurity to measure the quality of the splits, without specifying a maximum number of nodes. Three classifiers with these parameters were used, with 3, 5 and 7 estimators. The predicted class is decided by averaging the prediction probability given by each estimator. Because the RF classifier prediction is heavily dependent on the random selection of features used at each candidate split, 5-fold cross-validation was performed 50 times, shuffling the sample each time prior to fold attribution. In addition to that, due to the variability of the number of folds used by each author and the variability of their constitution, comparison between literature results and this study's results loses significance, therefore, LOO cross-validation was also performed 50 times, eliminating both of these issues.

The SVM classifier has two hyper-parameters parameters: C and γ_{SVM}. Because of the sheer volume of classifications to make, only the C parameter was tuned and the γ_{SVM} parameter was always considered $1/n_{feat}$ were n_{feat} is the number of features presented to the classifier. To choose the optimal C value, several (30-50) random splits of the original sample were made to establish as many train and test sets with a proportion of $0.8-0.2$ respectively. Then, instead of a grid search as it is usually done, a Python function was made to more efficiently search the C value that would optimize the macro averaged f-measure in the validation set. Optimization of the macro-averaged f-measure instead of the accuracy was motivated by the initial accuracy results of some FC matrices that were maximum when the classifier acted as a Most Frequent Class (MFC) classifier. To avoid these cases in which the classifier does not learn anything valuable, and to account for the differences in the classes' proportions, the macro-averaged f-measure was chosen as the measure to optimize as a function of C. Unlike RF classifiers, SVM classifiers are not based on random processes. However, to account for the previously mentioned variability in the fold attribution to each subject, 5-fold cross validation was performed 50 times and LOO cross-validation once.

After the definition of the classifiers and the model evaluation technique, a primary classification using each FC matrix individually, without feature selection, was made. Then, based on the performance of the classifier in the 5-fold cross validation, the methods of each family with the best macro-averaged f-measure were combined with each other in

groups of 2, 3, 4 and 5, making it a total of $13+{}^5C_2+{}^5C_3+{}^5C_4+{}^5C_5 = 39$ samples to classify. This approach was based on the similarity of the FC matrices calculated by methods of the same family and on the supposition that combining those methods would increase the amount of redundant information fed to the classifier. Macro-averaged f-measures were rounded to the third decimal digit and if two methods of the same family had equal values, the one with the better accuracy would be chosen. Finally, classification of the remaining 26 datasets with the aforementioned classifiers and evaluation techniques was performed.

The large number of features extracted per matrix poses a real problem to the classifiers. Combining matrices adds new information to the dataset that is going to be classified. Of this new information, some of it is expected to be redundant, some misleading and some different and beneficial, or as it shall be referred from here on, helpful. If most of this new information is helpful and the dimensionality of the data stays approximately the same, one expects a better classification performance than only with the original features. If most of the new information is misleading, then the classification is expected to be worse and if redundant, it isn't expected to change in a great manner. Thus, the classification performance with the added information is expected to depend on which of these three categories this new information mostly falls in. Also, in the cases where the dimensionality of the data greatly increases, the classifiers' behaviour becomes harder to predict due to the curse of dimensionality and the possibility of overfitting also increases. From the results of other authors in the ACPI database [93], most of the information is expected to be either redundant or misleading. Some classifiers deal better with this situation than others. RF classifiers, for instance, use only part of the total number of features for classification, making an automatic feature selection on the data in an attempt to use the best features of the whole set. In other classifiers, as the purely linear ones, a much bigger number of features affects more or less the classifier prediction depending on the shape of the classifier's decision function (a feature can also never affect the prediction of the classifier if the decision function is parallel to that feature's axis in the feature space). The latter is also the SVM classifier's case. If most of the information is indeed non-helpful, prior feature selection is needed to remove possible misleading features and to reduce the dimensionality of the data.

In this study, due to the probable need for a good feature selection, four different techniques were considered. First, the RF classifier's intrinsic feature selection was used. In a small subsample of the data, the results of this procedure with an SVM classifier, similar to the one described before, were not satisfactory and the method was left out. An iterative feature selection was thought of as a good alternative. As a consequence of the large number of features, a backward elimination approach would be infeasible, thus, a forward iterative feature selection was eventually attempted. The results of this technique were mixed and even with the forward selection approach the large number of features still made it difficult to employ this method (it needed approximately $13456 \times n_{iter} \times n_{mat}$ classifications, where n_{iter} is the number of iterations and n_{mat} is the number

of combined matrices). First, to counter this issue, a univariate feature selection was performed using an Analysis of Variance (ANOVA) test. The 1000 features with the lowest p-values were kept and the forward iterative feature selection was performed afterwards. The third considered technique was one developed by the author for the purpose of this study. In short, this technique consisted in classifying the subjects using every possible combination of two features and keeping in the final classification dataset the features that led to the n best results in the training set. This feature selection technique was performed after the previously mentioned univariate selection to reduce the classification time. Finally, after a univariate selection of the 500 features with the lowest p-values, a PCA representation of the features was obtained and the first two components were kept for classification. Of the last three techniques mentioned, after achieving similar primary results to the other two, the third was the chosen one since it was the one that enabled the completion of the whole classification process in the shorter period of time.

Even though the weights of all matrices ranged between 0 and 1, except in the case of the methods from the correlation family that ranged between -1 and 1, a data standardization was performed every time a new training set was defined by subtracting the training set average of each feature to the corresponding feature values of the whole data and dividing each value by the variance of the corresponding feature in the training set. Then, a PCA representation of the training data was obtained and the same transformation was applied to the test set avoiding any leakage of information. As stated before, the first two components were kept for classification. The whole classification process of the 39 datasets was repeated with this standardization and PCA based feature selection.

As a final note, it should be mentioned that the matrices from the undirected methods and the methods from the coherence family are symmetric, meaning that their calculation is commutative. In those cases, only one of the symmetric halves of the matrix was kept. For each pair of regions, the value given by the undirected counterparts of each family of methods is defined as the highest value in module of the two halves of the directed version's matrix, which can also be seen as a type of feature selection, hence the reason why this was only made in the classifications with prior selection of features.

4.4.2 Classification in the ADHD-200 datasets

Unlike the ACPI dataset, the ADHD-200 datasets have a predefined test set where the model evaluation is intended to be made. This removes the need for cross-validation since parameter tuning can be made in the predefined training set by randomly splitting it in a training subset and a validation set. An RF and an SVM classifier with the same characteristics as before were used with the same purpose of eliminating variables between the classification of the ACPI dataset and both ADHD-200 datasets. Both the filtered and the non-filtered datasets passed through the same procedures such that filtering would be the only variable. The steps to classification were analogue to the ones described for the ACPI dataset. There are, however, some subtleties that shall be mentioned. Since there is

now different validation and test sets, tuning of the C parameter was never made using information of the test set, removing any positive bias from the classification performance. The optimal C value found for the predefined training set was used in the classification of the predefined test set. Another consequence of this division of the original data is reflected in the choice of the methods to combine that can no longer be based on the results of the individual FC matrices in the test set, because that would introduce a positive bias in the results of the combined methods. The methods of each family to combine should be decided, thus, on a primary classification of the validation set. This was done 300 times in a new validation set randomly chosen each time from the predefined training set (20% of the sample). The methods of each family that originated the best macro averaged f-measures were combined in groups of 2, 3, 4 and 5, as was done in the ACPI dataset.

Classification with the RF classifiers of 3, 5 and 7 estimators was performed 100 times due to the already mentioned randomness associated with RFs. In the SVM classifier case, classification was performed only once with the optimal C parameter found in the training set.

Feature selection was also similar to the one performed in the ACPI dataset. After univariate selection of the 500 features with the lowest p-value and PCA transformation, an analysis of the first two components revealed that a higher number should be considered for the ADHD-200 datasets. As such, the first 80 components were kept for classification, instead of only two as was done for the ACPI dataset.

The general pipeline used from rs-fMRI to classification is summarized in Figure 4.1.

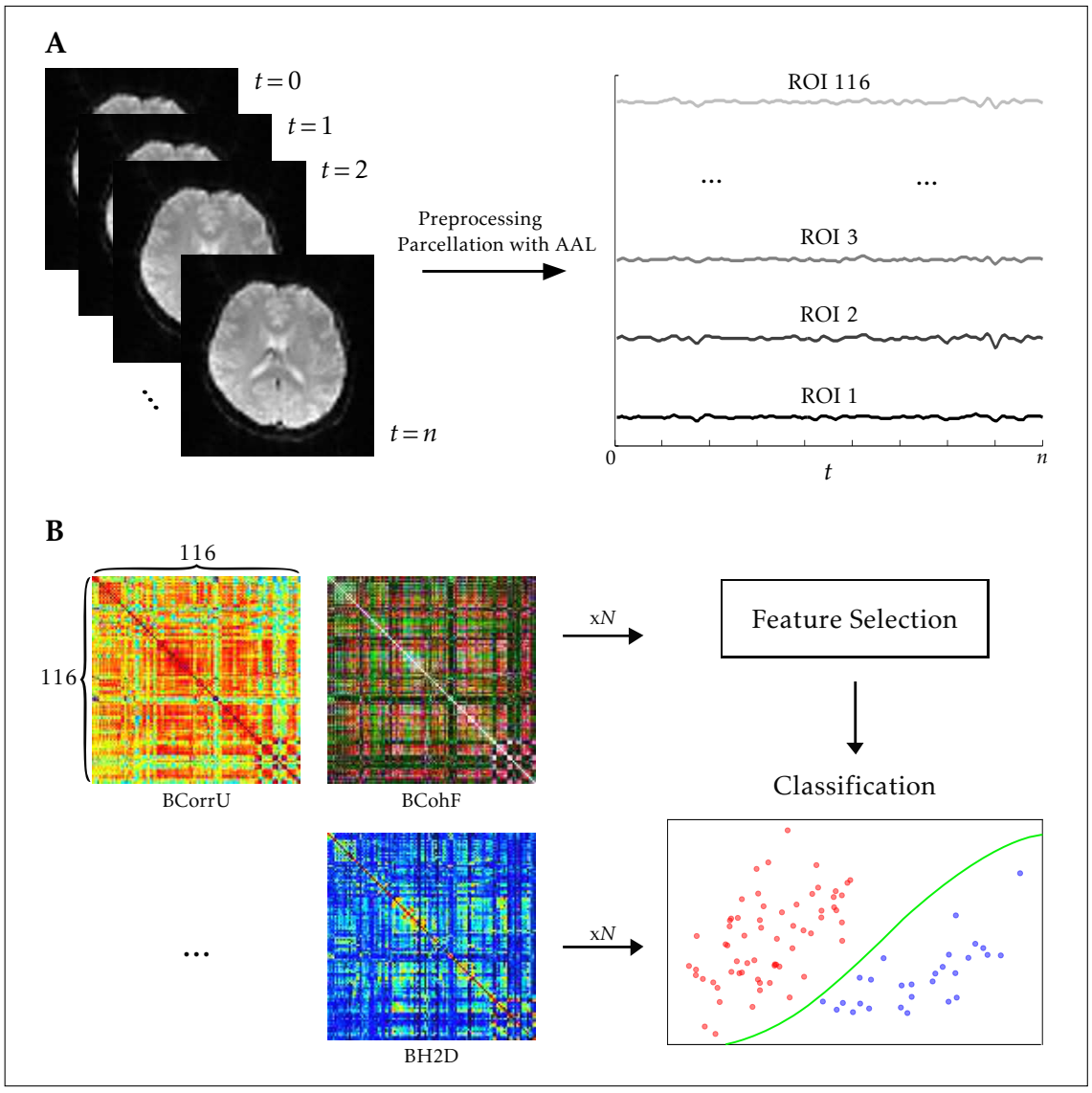


Figure 4.1: From rs-fMRI to classification.
(A) BOLD signal time series (n time-points) is acquired for each voxel (here only a slice is shown). After preprocessing and parcellation with AAL the time series of each ROIs' voxel is averaged (116 ROIs).
(B) Each statistical method is used to calculate the corresponding FC rs-fMRI matrix. The process is repeated N times where N is the number of subjects of the whole dataset, such that there is one matrix of each method for every subject. The weights of the matrices are used as features for classification, with and without prior selection of these features.

Chapter 5

Feature Set and Classifier Analysis

Before making a given classification, it is always useful to first try to visualize the data to be handled by the classifiers, i.e. the feature set, for one to identify possible differences between the classes and to get a sense of how well the classifiers will be able to generalize to the test set. Also, the classifiers should be tested in simple cases to assess if they are well constructed and if the classification pipeline to be used works properly.

Here, two classes are going to be considered separable if one can find a set of rules based on a given number of features that would allow a perfect labelling of every data point or instance. An example of two separable classes are the adult-by-law class and the children-by-law class. Based on the age and nationality (features) of the subjects (instances) one can perfectly separate the two classes using a rule (the age of majority in a given country).

As already stated, supervised ML algorithms use the information present in a given subset of instances or data points, the training set, to define the rules that best separate two or more classes. Classes can be separable in the training set, but not in all datasets one can form using instances of those classes. For instance, in the previous example, if the training set is the French population, the classifier might define that every instance with feature-age value greater or equal than 18 belongs to the adult class, effectively separating the two classes in the training set since the age of majority in France is eighteen. However, the adult and the children classes are not separable using only the feature corresponding to the age of the subject. In Scotland, for instance, the age of majority is sixteen, which means that the classifier would attribute all seventeen-year-old Scots to the children class when in reality they belong to the adult class.

In this text, we shall refer to class separability in a subjective way. For instance, two easily separable classes will be considered classes that would be perfectly classified by most classifiers and two non-separable classes as two that would not be separable with

any classifier and, in fact, by any set of classification rules. Also, one could refer to class separability in a given subset of instances, in all existing instances or to the separability in a sample of infinite size. Two classes are easily separable in a given dataset if the number of features per subject is greater than the number of instances of that dataset. This is another problem with having a high number of features relative to the sample size. If one has more features than instances, then the performance in the training set is rarely less than perfect, which means that parameter optimization has to be made judging only the performance on the validation set. In the next Section (Section 5.1) we shall briefly investigate the feature sets used in this study in terms of how they allow one to separate the ADHD and the HC classes. One would like to assess the feature set and the separability of the two classes without introducing the variables associated with the good or bad quality of the classifiers, something to be addressed only in Section 5.2.

5.1 Feature set

Several classifications are going to be made in this study, the feature set to be used for each of those classifications will always be a subset of the feature set composed by the weights of all FC matrices. If one looks at the mostly poor classification performances already reported in both databases used in this study (see Chapter 3), one would guess that most features extracted from the FC calculated with the statistical methods used in them, mostly correlation based methods, are not very useful for classification, since if the opposite was true, the classification performances would be far better than the ones reported. However, apart from distinct statistical methods, different pipelines to calculate each subject's FC from the ones used in this study were sometimes adopted, which makes this assumption not a certain one.

Classifications based on a small number of features tend to generalize better than classifications based on a big number of features because it is less probable for the rules defined with a small feature set to be specific to the training set. Thus, to assess the separability of the ADHD and the HC classes in the world population, one would like to look at their separability in a small subset of the original feature set. As stated in Chapter 4, an ANOVA test is going to be used for univariate feature selection. However, since there are so many features in analysis, it can happen that the two classes yield a low probability of belonging to the same distribution just by chance, which would make the two classes almost separable in the training set but not separable in a new dataset of the same distribution. What could one conclude if the two classes were still somewhat similar in the feature with the lowest p-value under the null hypothesis that the means of the two distributions are equal? This question does not have an easy answer because even with the two classes having the same distribution in all features, the relationship between the features of a given subset could be enough to separate them. However, if the two classes have different distributions in one feature, just that is already enough to almost separate them, even without looking at any other relationships. Thus, as a very simple

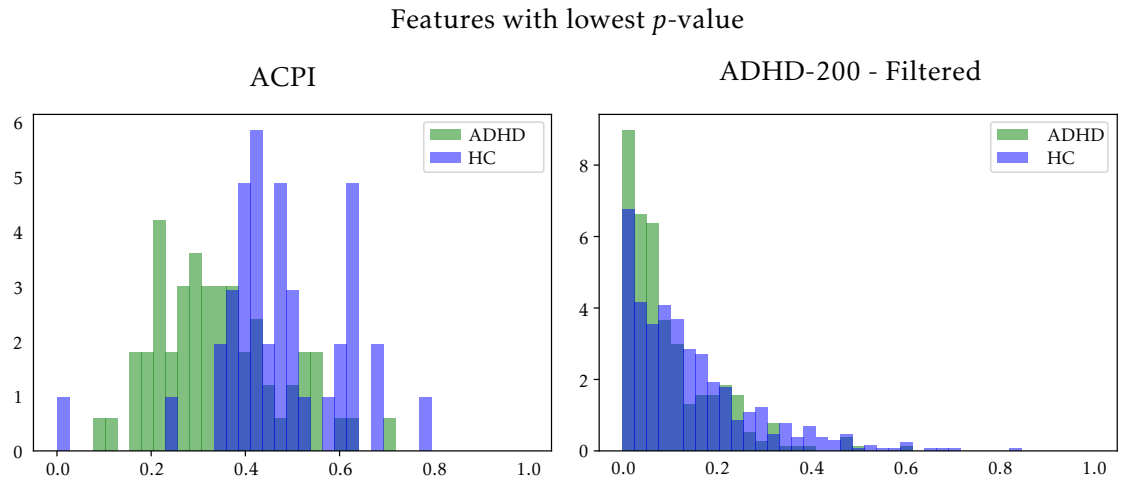


Figure 5.1: Normalized distribution of the ADHD and the HC classes in the feature with the lowest p-value from a binary ANOVA test in the whole ACPI dataset (left) and in the whole (train set plus predefined test set) ADHD-200 dataset (right). The feature on the left corresponds to the wavelet based coherence between AAL regions 2322 and 2111; p-value: 1.8×10^{-5}. The feature on the right corresponds to the Fourier based coherence for 0.43 Hz between AAL regions 9160 and 6222; p-value: 4.8×10^{-7}. In the non filtered ADHD-200 dataset (not shown) the feature with the lowest p-value corresponds to the wavelet based coherence between regions 7101 and 7011; p-value: 5.1×10^{-7}. Note: superposition of the two histograms appears in dark blue.

way to assess the separability of the two classes, one could look what the lower bound of this separability is expected to be in the whole set by looking at their distributions in the feature with the lowest p-value in a binary ANOVA test. This is shown for the ACPI and the whole ADHD-200 datasets in Figure 5.1.

As it can be seen, in the ACPI dataset the two classes are almost separable using only the feature shown in Figure 5.1. However, in the ADHD-200 filtered dataset, the two classes significantly overlap, and not much can be said about their separability using the depicted distributions. Just by looking at these two features, one could say that the ADHD and the HC classes seem easier to classify in the ACPI dataset than in the ADHD-200 datasets. Additionally, one should have in mind that even though by using these features the classifier would be able to improve the classification performance relative to the random case, the distributions of both classes in these features in the training set have to allow the classifier to predict their importance for classification in the test set. Thus, a prediction of how well the classifier is going to perform in either database is not as easy as looking at the distributions of both classes in these two features in the whole dataset.

For one to be able to refer to the usefulness of the two features shown in Figure 5.1 in the classification of subjects in databases outside the ones where their selection was based, one cannot rely on the p-values yielded by the ANOVA test. There are several reason as to why this is true. The first one is that even though the ANOVA test is robust against non-normal random variable distributions, the validity of this and other assumptions made by this statistical test would have to be assessed in order for this analysis to make sense. Besides that, due to other variables such as the acquisition machines, the machine

operators, the acquisition protocols and the preprocessing pipelines, the distributions of the two classes in both of these features in new datasets could be completely different than the ones from where their values were drawn in the databases used in this study.

Besides their benefits to classification, dimensionality reduction and feature selection techniques can be used to derive low-dimensional representations of data that would allow one to visualize it. In the classifications that used features extracted from the ACPI dataset, the ones with prior feature selection used only two features. This allows one to go one step further from the previous analysis and see exactly what the classifier will be trying to classify. As an example, let us randomly split the ACPI database in a training and a test set 0.8-0.2 as in a 5-fold cross validation. After selecting the 500 features of the training set with the lowest p-values and plotting the two principal components of their PCA representation against each other, the result is what is shown in Figure 5.2.

The two classes are almost linearly separable in the training set using only the data's two principal components[1]. In the test set the same scaling and PCA transformations applied to the training set yield two much more overlapping distributions, which means that, in some of the features selected with the ANOVA test, the two classes did not carry the same difference between their distributions' means to the test set. This behaviour was verified to be the rule rather than the exception by repeating the process of randomly splitting the dataset in two other proportionally analogue parts and visualizing the two principal components of the training and the test set after the same transformations. From the results of this experience, one expects the classifier to generalize little above the random prediction mark in the ACPI dataset.

A possibility to improve the classes' separability in the test set is to change the number of features selected with the univariate tests or the number of principal components (though an increase of the latter will result in a feature set much harder to visualise). The problem in doing that is the resulting overestimation of the classifier performance that comes with using information from the test set to choose the type of feature selection applied to the data.

In short, the ADHD and the HC classes seem to be similar in both databases and the univariate feature selection could not be enough to counter the problems resulting from the high-dimensional feature set used. Other techniques to visualize the data and the possible generalisability of the classifier could be used such as the t-SNE [142] algorithm and other manifold learning algorithms.

5.2 Classifier

Before using the classifiers in classes where the decision function that separates them is not known, one should first analyse if they are working properly. To do this, a Python function was made to simulate two features in which the distributions of the ADHD and

[1] After this visualization test the number of components to feed to the classifier in the ACPI classifications was chosen as only two to avoid overfitting the training set.

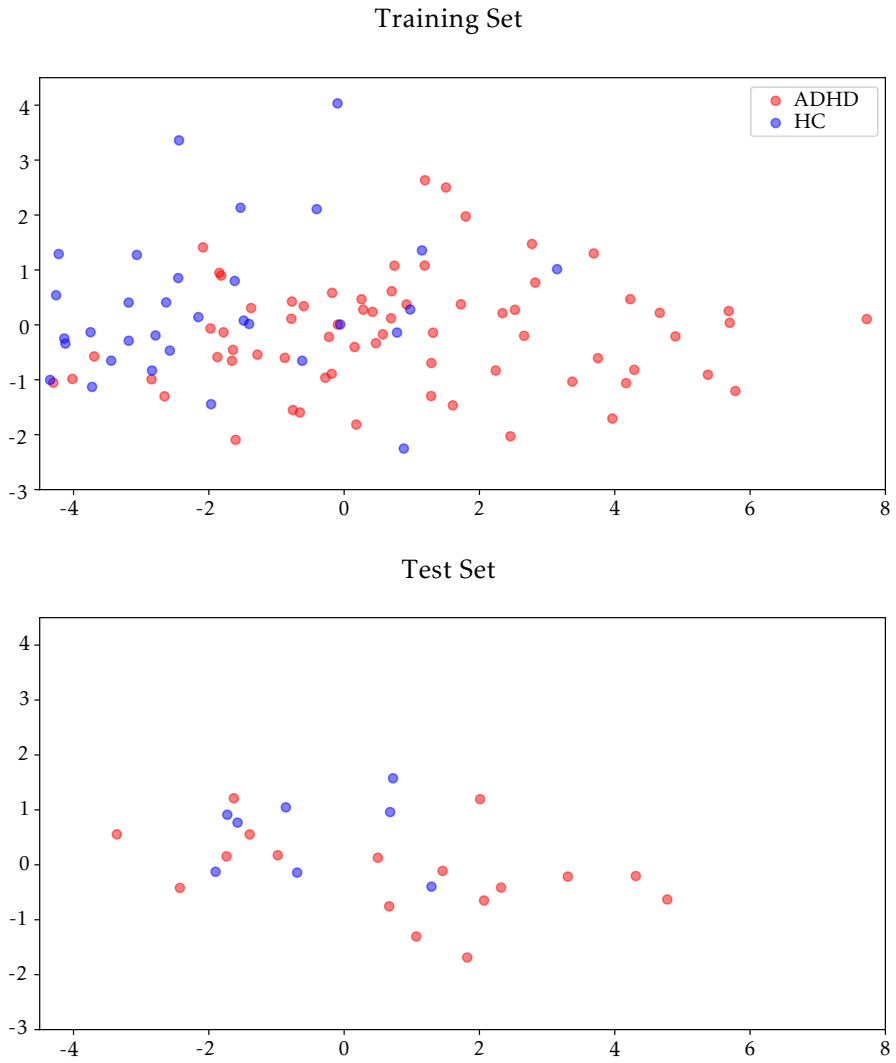


Figure 5.2: Distribution of the ADHD and the HC classes in a randomly selected training (top) and test (bottom) sets from the ACPI dataset (80-20), extracting features from the directed correlation matrix. Horizontal axis: principal component of the 500 features with the lowest p-value of an ANOVA test. Vertical axis: second component.

the HC classes in the feature set are separable by a previously given polynomial function of the type $y = a_1x + a_2x^2 + ... + a_nx^n + b_0$, $x \in [0,1[$, $y \in [0,1[$. The data points of each class were distributed randomly inside their corresponding fraction of the feature set. This was made for the subjects in the ACPI dataset and a 0.8-0.2 random split was made twenty times to make a training and a test set respectively, using a linear, a quadratic and a cubic boundary between the two classes. SVM and RF classifiers analogues to the ones described in Chapter 4, achieved, in the twenty repetitions, an average accuracy above 0.95 in the test set, using only these two simulated features, for the linear, the quadratic and the cubic separation cases. Parameter optimization was made in the test set. These results prove that the classifiers can successfully define generalization rules using the training set.

An interesting analysis that would allow one to evaluate the quality of the classifiers

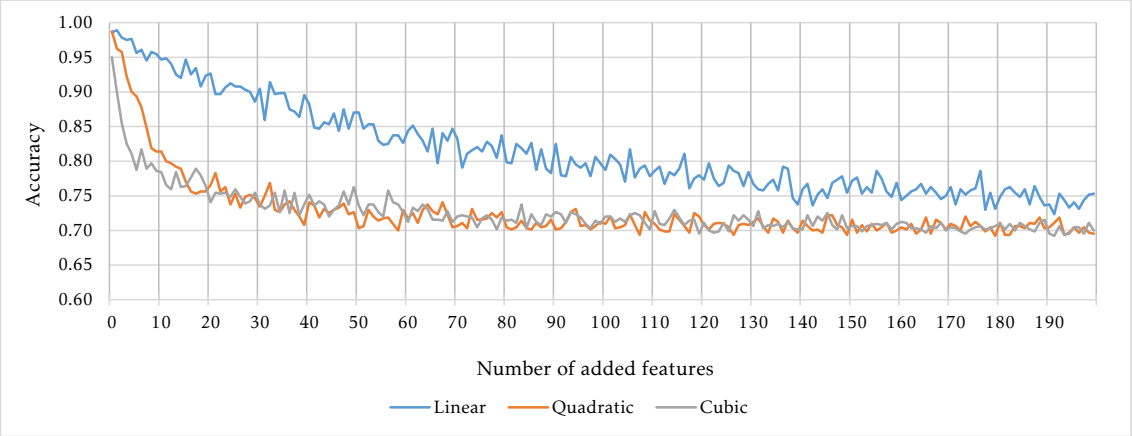


Figure 5.3: Average accuracy of an SVM classifier as a function of the number of added features (up to 200) to a feature set composed of two simulated features separable with a linear (blue), a quadratic (orange) and a cubic (grey) decision function. Linear decision function: $y = x$, quadratic decision function: $y = -4x^2 + 4x$, cubic decision function: $y = 16x^3 - 24x^2 + 9x$, where x and y are the two simulated features.

while analysing the feature set at the same time, would be to plot the classifier performances in the test set as random features from the original feature set are added to the two simulated features. This plot is shown in Figure 5.3 using an SVM classifier also analogue to the ones described in Chapter 4.

When no original features were added to the two simulated features, the average accuracy in the test set over twenty repetitions was above 0.95 when the boundary between the two classes was linear, quadratic or cubic, as previously said (Figure 5.3). The difference in accuracy between the case where only the two simulated features were used for classification and the case where those two features plus ten others randomly chosen from the original feature set were used, is not much when the boundary between the two classes in the simulated features is linear. However, the decrease in performance from zero to ten added features, when the boundary was quadratic or cubic was very accentuated, passing from 0.99 to 0.81 in the former case and from 0.95 to 0.78 in the latter. With only just ten original features added to the simulated features the classification in the test set dropped almost 0.2 in average accuracy. This is a very important result. Even if there are two features in the original feature set that enable a perfect classification with a decision boundary as simple as a quadratic function, the benefit of this relationship is dramatically absorbed by even just a small number of other added features from the original set. Until two hundred added features, the average accuracy steadily decreased in all cases, especially in the linear one, to just above the MFC classifier mark of 0.68. With an RF classifier, results are slightly better because classification is based only on the most discriminant features of a given subset, making the effect of adding features less noticeable.

Comparing the results shown in Figure 5.3 to the same experiment but adding features with random values for either classes with a uniform distribution, one notices that only in

the case where the boundary between the two classes is linear in the first two simulated features, that adding features from the original feature set is significantly better than adding features with random values. This further confirms the lack of utility of most features for classification of the ADHD and the HC classes.

To investigate the extent to which the classifiers are affected by adding features randomly chosen from the original feature set to features that benefit classification, an experiment was made where a feature was simulated, in which all subjects of one class had the value one and the subjects of the other class had the value zero. Then, features were added to the simulated one as in the previous experiment. An SVM classifier was found to drop from perfect classification in the test set when 198 features were added. As an example, when all 13456 features from the undirected bivariate correlation matrix were added, results were similar to the case where the simulated feature was not present and only the features from this matrix were used. These results show the importance feature selection might have in the performance of the final classifications, if such discriminant relationships exist within a small subset of features of the original feature set.

Chapter 6

Results

In this work, five variations were introduced in a standard procedure of subject classification to an ADHD class and a control class[1], using data from two databases: ACPI [1] and ADHD-200 [4, 15]. As explained in the methods, the pipeline from rs-fMRI data to classification was basically: data acquisition → pre-processing → construction of FC and EC matrices → feature selection → classification. All but the first two steps were performed by the author. The five mentioned variables were introduced in several points of this pipeline: 1) filtering or no filtering of the data in the preprocessing step; 2) the method used to calculate the FC and the EC; 3) the number of matrices used to construct the feature set; 4) the use of feature selection or not; and 5) classification using SVMs or RFs. Each classification was made following the same steps and given the same opportunities to remove any possible bias towards a particular statistical method or result in general.

Accuracy, macro-averaged f-measure, TPR and precision measures of all classification variants performed in the course of this study are reported in Annex I of this text, along with the accuracy value achieved with an MFC classifier in the same conditions (MFC value), the accuracy value achieved by a random classifier that takes into account the classes' balance (WRP_acc) and the macro-averaged f-measure achieved with a coin-toss like classifier (RP_fm). Also, note that the classification results using RFs reported for each method are the best ones out of the three made with 3, 5 and 7 estimators.

[1] The positive class was always considered the ADHD class and the negative class was always considered the HC class.

Table 6.1: Used methods and their notation.

Methods	**Code**
Undirected Correlation	BCorrU
Directed Correlation	BCorrD
Fourier Coherence*	BCohF
Wavelet Coherence	BCohW
Undirected h^2	BH2U
Directed h^2	BH2D
Undirected Mutual Information	BMITU
Directed Mutual Information	BMITD1
Directed Mutual Information (corrected)	BMITD2
Undirected Transfer Entropy	BTEU
Directed Transfer Entropy	BTED

*BCohF composed of BCohF[0]: 0.01 Hz , BCohF[1]: 0.043 Hz and BCohF[2]: 0.076 Hz.

6.1 Comparison of Methods

Thirteen matrices were constructed, all but three with a different method. These three resulted from the Fourier based coherence at three frequencies 0.01 Hz, 0.043 Hz and 0.076 Hz. The Fourier based coherence in these three frequencies will be denoted as BCohF[0], BCohF[1] and BCohF[2], respectively. All other methods will follow the notation introduced in Wang's paper [147] as shown in table 6.1.

Figure 6.1 shows the average accuracy and macro-averaged f-measure obtained in a LOO cross-validation with all thirteen matrices individually in the ACPI dataset. Figure 6.2 shows the values of these same measures but in the test set of the ADHD-200 filtered dataset. In the two Figures, probably the most obvious difference is how much better the SVM classifier's performance was. The RF classifier performed just above the coin-toss like classifier when evaluating the macro-averaged f-measure in the ACPI dataset and performed worse than that in the ADHD-200 filtered dataset. Moreover, in the ACPI dataset, the mutual information family methods performed much better than every other when using the SVM classifier, actually surpassing the only classification performance reported to date using the ACPI database to separate subjects with ADHD of HC [93]. In that study, a maximum macro-averaged f-measure of 0.60 was achieved using a LASSO classifier and a LOO cross-validation to evaluate the model. Here, a macro-averaged f-measure of 0.677 was achieved with an accuracy of 0.744 using the BMITD2 method and selection of features as described in Chapter 4. Interestingly, in the ADHD-200 filtered dataset, and particularly when the SVM classifier was used, the methods from the mutual information family performed poorly, both in terms of macro-averaged f-measure and in terms of accuracy. Though, the opposite happened in the validation sets, before

classification of the test set. This suggests that the SVM classifier might have overfitted the training set, an hypothesis backed by the value of C found by the algorithm, which was the maximum possible in all three mutual information based methods. Thus, the results achieved with mutual information based methods in the filtered test set should be evaluated in a conservative way.

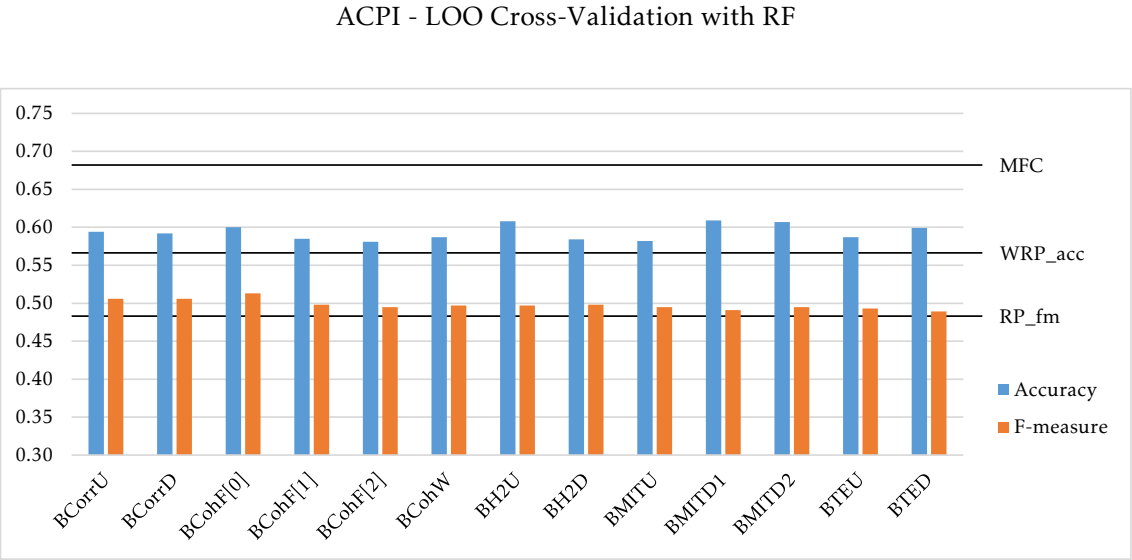
Correlation and transfer entropy based methods performed generally well. The latter especially in the ADHD-200 sample. The poor result achieved with BTED in the ACPI dataset could be a consequence of underfitting because accuracy values close to those achieved with an MFC classifier, combined with low macro-averaged f-measure values usually indicate that. Coherence and h^2 based methods performed poorly in the ACPI dataset. In the ADHD-200 filtered dataset, Fourier based coherence at 0.43 Hz was the best method not only with the SVM classifier but also using RFs, which reassures the importance of this method in this case. Though, the same did not happen with the other coherence family methods.

Regarding the individual matrices, the directed and undirected variants of the methods seemed to achieve on average the same results. For instance, classifications with the SVM classifier using directed methods were, on average, 0.6% worse in the ADHD-200 filtered dataset and 1.0% better in the ACPI dataset.

6.2 Feature Selection vs No Feature Selection

The results obtained with and without feature selection using an SVM classifier are reported in Annex I as well. Figure 6.3 shows how feature selection affected classification accuracy and macro-averaged f-measure. As it can be seen, on average, the same classification with and without feature selection does not have an abruptly different value of macro-averaged f-measure, either in the ACPI dataset or in the ADHD-200 filtered test set. Though, the standard deviation of the average of this change (not shown in Figure 6.3) is of 5.0% and 11.7% in the ACPI and the ADHD-200 test set, respectively. These values indicate that one should be careful when evaluating the effect of feature selection in the performed classification, just by taking into account the average of this effect. For instance, if one looks at the maximum increase of macro-averaged f-measure to the selected case in Figure 6.3, he sees that it reached 17.34% in the ADHD-200 sample, when using the BH2D method. Regarding accuracy values, there is also an opposite trend in the datasets from the ACPI and the ADHD-200 sample, though, more pronounced than in the macro-averaged f-measure case.

In spite of being a rather simplistic summary of what happened in all classifications when feature selection was applied, Figure 6.3 allows one to say that, generally, feature selection tended to improve generalization ability in the ADHD-200 filtered dataset. In the ACPI dataset, however, the opposite cannot be as easily said. First, because the measure that was optimized and that interests us more is the macro-averaged f-measure and the



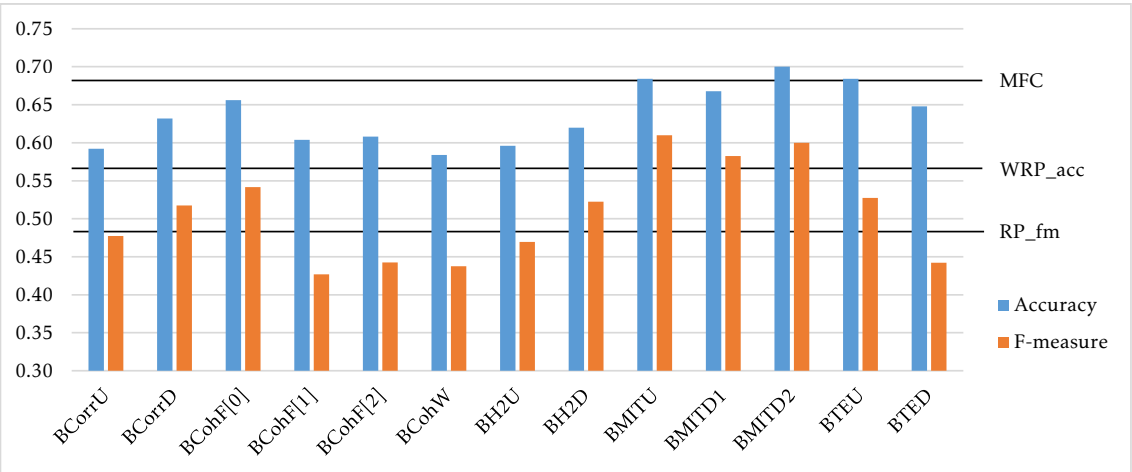
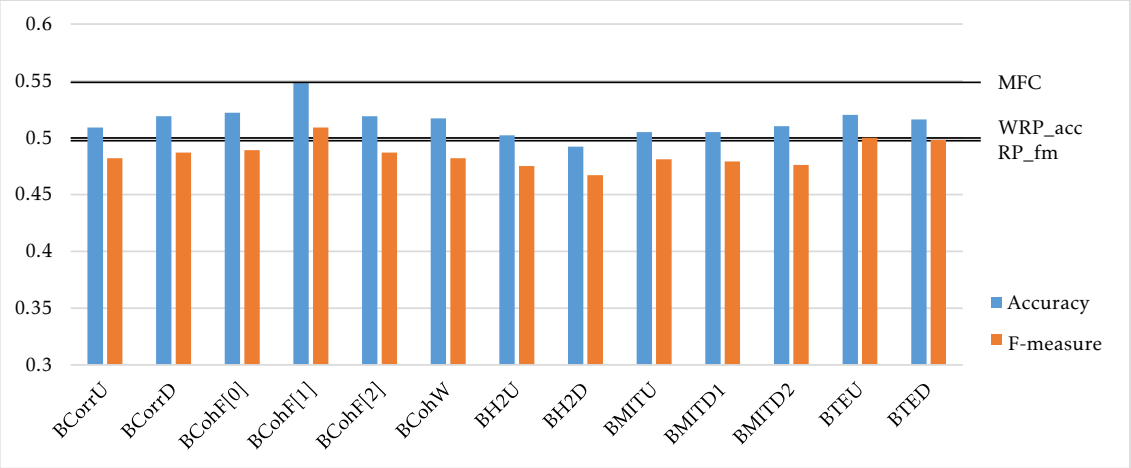


Figure 6.1: Average accuracy and macro-averaged f-measure of all 50 LOO cross-validations using an RF classifier and of both LOO cross-validations using an SVM classifier with and without feature selection in the ACPI dataset. MFC line, MFC classifier accuracy (0.68); RP_fm line, average macro-averaged f-measure of a coin-toss like classifier performed 50 times (0.482); WRP_acc line, average accuracy of a random classifier that takes into account the class proportions performed 50 times (0.565).

ADHD-200 RF in Predefined Test Set



ADHD-200 SVM in Predefined Test Set

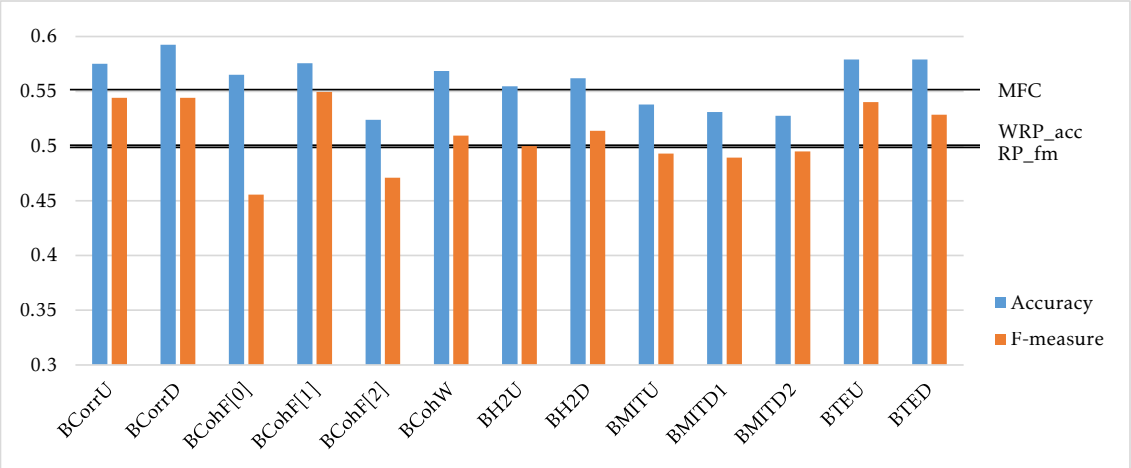


Figure 6.2: Average accuracy and macro-averaged f-measure of all 100 classifications using an RF classifier and of both classifications using an SVM classifier with and without feature selection in the ADHD-200 predefined test sets. MFC line, MFC classifier accuracy (0.548); RP_fm line, average macro-averaged f-measure of a coin-toss like classifier performed 100 times (0.498); WRP_acc line, average accuracy of a random classifier that takes into account the class proportions performed 100 times (0.504).

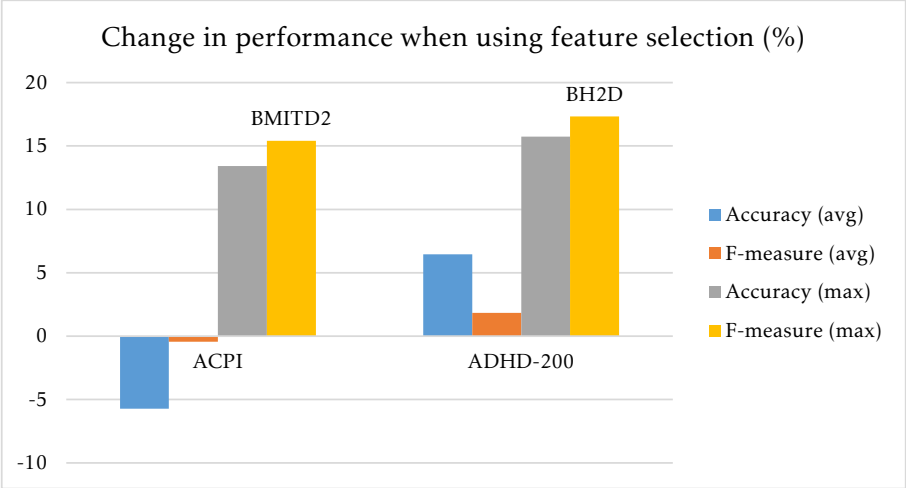


Figure 6.3: The accuracy and macro-averaged f-measure values of all similar classifications in the ACPI and the ADHD-200 filtered test set, were averaged with and without feature selection and the respective change was calculated in respect to the values obtained when no feature selection was used (Accuracy (avg) and F-measure (avg) bars). The maximum values of this change are also reported (Accuracy (max) and F-measure (max) bars). Average accuracy and macro-averaged f-measure standard deviation was 5.2% and 5.0% respectively in the ACPI dataset and 6.3% and 11.7% in the ADHD-200 filtered test set. Notes: 1) when classifiers acted as an MFC classifier, those values and the respective values when feature selection/no feature selection was used, were not included in the average or maximum calculations. 2) values obtained in the ADHD-200 sample when various connectivity matrices were combined were not included in the calculations of the relative changes in the classifier's performance because different matrices were combined when feature selection was used and when it was not.

relative change in such measure was $> -1\%$, being actually positive if only LOO cross-validation measures are considered. Secondly, even though the average classification in the ACPI dataset was lower, the best classification out of the two cases was achieved when feature selection was made. This means that feature selection was helpful to some methods in the ACPI dataset.

Finally, it should be noted that on several occasions the SVM classifier performed as an MFC classifier when feature selection was used and none when it was not used. This result shall be discussed in Chapter 7.

6.3 Filtering vs No Filtering

Two preprocessed datasets sampled from the ADHD-200 database were used, in one a bandpass filter 0.009-0.08 Hz was applied to each time series and in the other one not. Classification in these datasets was made with RFs, and also with and without feature selection using SVMs. The overall impact filtering had on classification is summarized in Figure 6.4, which represents the average change in classification accuracy and macro-averaged f-measure when filtering had been made relative to when it had not been made. When RFs were used to classify the ADHD-200 filtered test set, classifications were, on average, as accurate and had slightly better macro-averaged f-measure than when

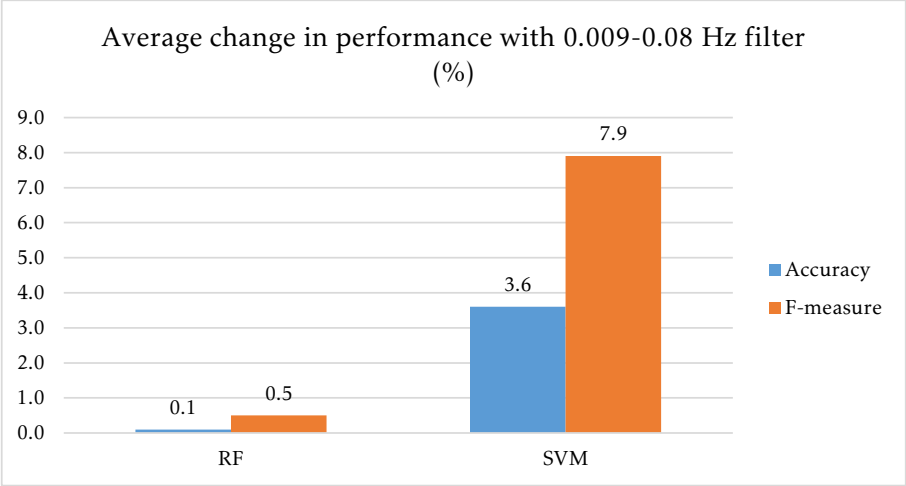


Figure 6.4: Average change in classification performance when ADHD-200 data preprocessed with a 0.009-0.08 Hz filter was used, relative to the classification performance when unfiltered data was used (values in percentage). Values shown for both classifiers. Notes: 1) when classifiers acted as an MFC classifier, both filtered and non filtered values were not included. 2) values when various connectivity matrices were combined were not included because different matrices were combined in the filtered and non filtered cases.

they were used in the non filtered dataset. When SVMs were used, the difference was significantly greater. A 7.9% increase in macro-averaged f-measure was, on average, verified in this case. It can be concluded, thus, that filtering had a positive impact on generalization ability, especially when feature selection was used, reaching an average increase in macro-averaged f-measure of 10.2%. Also, maximum accuracy and macro-average f-measure results were always achieved when using filtered data. To be noted that of all coherence based methods, BCohF[2] seemed to be consistently less benefited by filtering.

Interestingly, when validating the parameters and deciding which matrices to combine (Tables I.7, I.9, I.11, I.13, I.15 and I.17 of Annex I) filtered data did not have a big impact on classification accuracy and macro-averaged f-measure, with an average negative change $> -1.5\%$ in accuracy relative to the case where no filter had been applied, when using both RFs and SVMs, and -0.9% in macro averaged f-measure for RFs and 1.4% for SVMs.

6.4 Single Matrix vs Multiple Matrices

The main goal of this study was to evaluate whether using connectivity information from several statistical methods at once would result in better classification performances than using information from just one. As already noted when looking at the classifications' results with a single connectivity matrix in Section 6.1, the SVM classifiers generally performed much better than RFs, thus, we shall focus here primarily on the results of the former. The macro-averaged f-measure distribution of the classifications using a single

matrix or a combination of matrices on the datasets of both databases is shown in Figure 6.5. In the ACPI dataset, when there was no prior feature selection to classification (top-left corner of Figure 6.5), combining matrices was not helpful. The average macro-averaged f-measure was 0.506 when a single matrix was used and 0.484 when multiple matrices were used. These results are not particularly good, since they are around 0.5, which is the value a random classifier would have when taking into account the proportion of each class and just above the performance of a coin-toss like classifier. However, about 40% of the classifications in this context were above the 0.52 mark (top-left corner of Figure 6.5) when a single matrix was used and only about 8% when multiple matrices were combined. Also, the good performances achieved with mutual information based methods were not carried on when other matrices were combined with the best method of this family, the BMITU, in 5-fold cross-validation. Though, the best results when two matrices were combined always had the BMITU matrix as one of them. When feature selection was made (top-right corner of Figure 6.5) there was not as big a difference in the distribution of the results with combined matrices as there was in the distribution of classifications that used only a single matrix. Though, a higher number of classifications in both cases were above the coin-toss like classifier threshold.

In the ADHD-200 test sets, results are also poor in general, though, the four best classifications surpassed the 0.61 accuracy mark of the best classification in the ADHD-200 competition [44], as far as the publicly available test set is concerned. Those classifications were all obtained with a combination of matrices and using feature selection, one, using connectivity matrices built from the BCohF[1] and the BMITD2 methods, achieved 0.589 of macro-averaged f-measure and an accuracy of 0.616. The other three were obtained in the non filtered ADHD-200 test set (results not reported in Figure 6.5) and achieved: 1) 0.566 of macro-averaged f-measure and 0.623 of accuracy, using the methods BH2U and BMITD2; 2) 0.623 of macro-averaged f-measure and 0.658 of accuracy, using the methods BH2U and BTED; and 3) 0.648 of macro-averaged f-measure and an accuracy of 0.678, using the methods BH2U, BMITD2 and BTED.

When no feature selection was made, the distribution of macro-averaged f-measure values in the ADHD-200 filtered test set (bottom-left corner of Figure 6.5) was very similar in the single matrix case and in the case with multiple matrices. The average of both distributions is, again, close to 0.5, being 0.507 in the single case and 0.504 in the combined case, and a slightly more pronounced difference is reflected in the average accuracy, 0.544 in the single case and 0.561 in the multiple case. Though, it seems that three methods, BCorrD, BMITD2 and BTEU, were consistently helpful for classification when combined with other methods and with themselves (see table I.16), being the highest accuracy of this category achieved when the three were combined, 0.603, and also reaching a macro-averaged f-measure of 0.563, at a mere 0.002 of the best in this dataset. In the non filtered ADHD-200 test set, similarly to what happened in the ACPI dataset but not in such a pronounced way, mutual information based methods originated the best accuracy and macro-averaged f-measure values.

Figure 6.5: Top-left corner, normalized macro-averaged f-measure distribution in the ACPI dataset without feature selection. Top-right corner, normalized macro-averaged f-measure distribution in the ACPI dataset with feature selection. Bottom-left corner, normalized macro-averaged f-measure distribution in the filtered ADHD-200 test set without feature selection. Bottom-right corner, normalized macro-averaged f-measure distribution in the filtered ADHD-200 test set with feature selection. In orange are represented the values obtained when a single matrix was used for classification, in blue are represented the values obtained when multiple matrices were used for classification. Notes: 1) values of both LOO and 5-fold cross-validations were included in the histogram, though only the average value for each method in the 5-fold case was considered. 2) values where the macro-averaged f-measure was not defined, were not included.

When feature selection was made, more interesting results were achieved. In this case, both distributions are very oddly shaped, especially when compared to the case where no feature selection was used. Both are characterized by a reasonable number of very high and very low macro-averaged f-measure values. Some of the best methods in the classifications without feature selection acted as an MFC classifier once feature selection was made and also, the ones that did not, usually did not carry their good results to the corresponding selected case. This latter characteristic was also true for the worst methods. To measure this phenomenon, one can use, for instance, the standard deviation of the percent change in macro-averaged f-measure, when feature selection was introduced, relative to the values obtained when no selection was made (the same type of change calculated in Sections 6.2 and 6.3). This standard deviation is 15.4% while the average is 0.5% which reflects how much one method increased or decreased in the corresponding selected or non selected case. The method BCohF[1] seemed to be the one that most benefited classification when combined with others.

It is also interesting to look at how affected was a given method when other matrices were combined with it (Figure 6.6). By giving a general look at the four graphs in Figure 6.6 one immediately notices how all seem to point downwards, meaning that combining

the best 5 methods from each family was never better that using the best method alone. A closer look reveals that this also happened not only when 4 matrices were added to the best statistical method but also when any other number of matrices were added to it. In the ACPI dataset without feature selection (top-left corner of Figure 6.6) except the worst method, which was the BTED, no method benefited, on average, from the combination with another method. With feature selection (top-right corner of Figure 6.6), the results were similar but better when 4 matrices were combined than when only 3 matrices were combined. In the ADHD-200 filtered test set without feature selection (bottom-left corner of Figure 6.6) the two linear methods, BCorrD and BCohW, that lead to a healthy margin to the random classifier mark, did not benefit, on average, from their combination with any number of matrices. Though, the second best accuracy and a good macro-averaged f-measure mark of 0.558 (table I.16) was achieved when they were both combined. The BMITD2 matrix was the only one that consistently benefited from the combination with other matrices in the ADHD-200 filtered test set without feature selection. However, as already said, the result achieved using only this matrix has to be cautiously considered.

When feature selection was made (bottom-right corner of Figure 6.6), not as many results are available due to the already referred cases in which the classifier acted as an MFC classifier. Though, this was the only context where, on average, a matrix combined with other had a better classification performance than the best performance achieved using only a single matrix (BCohF[1] when combined with a single matrix). The steep drop in performance when the three matrices were combined with the one from the BCohF[1] method comes partially from the fact that the classifier, when all methods but the one from the transfer entropy family were combined, acted almost as an MFC classifier because the precision was 1 and the TPR was close to zero, which means that there were no false positives and many false negatives, leading one to conclude that probably most subjects were attributed to the negative class. This combination had a macro-averaged f-measure of 0.370, which clearly had a big impact on the average value achieved when three matrices were combined with any of these three. Finally, it should be noted that the best results achieved when multiple matrices were combined, came mostly when feature selection was used.

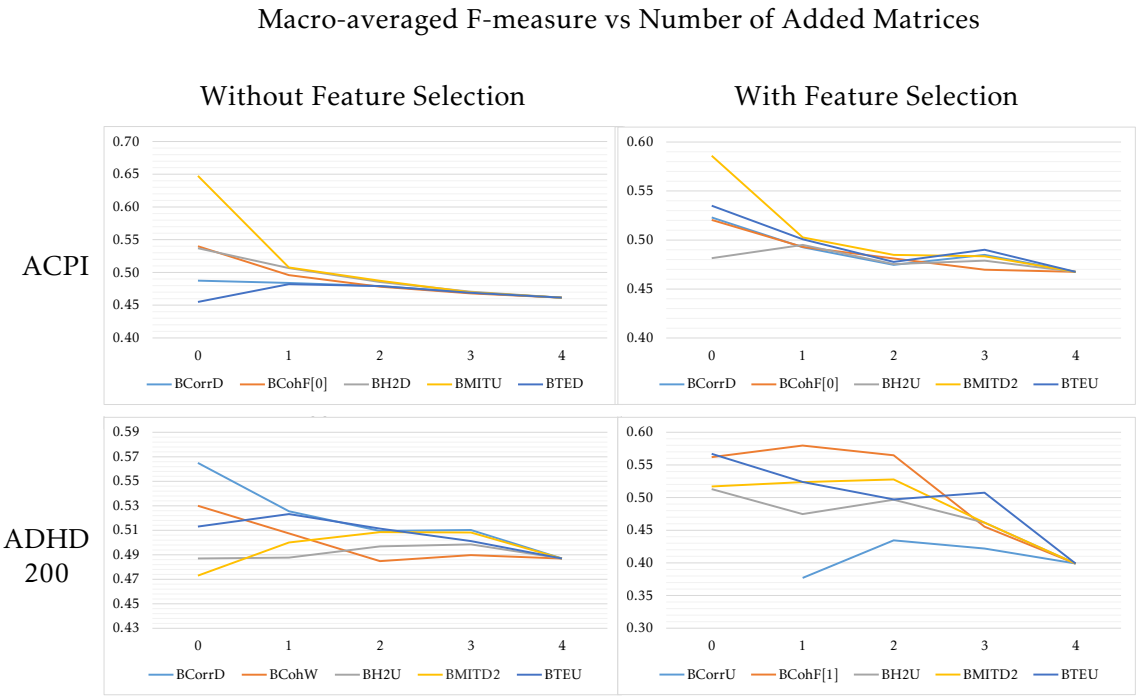


Figure 6.6: Average macro-averaged f-measure (vertical axis) of classifications using SVMs, when 0, 1, 2, 3 and 4 matrices (horizontal axis) were added to the method referred by each line, in the ACPI dataset without and with feature selection (top half) and in the ADHD-200 filtered test set (bottom half). Light blue, best correlation family method; Orange, best coherence family method; Grey, best h^2 family method; Yellow, best mutual information family method; Dark blue, best transfer entropy method.

Chapter 7

Discussion

As expected from the results already reported by other authors (see Chapter 3 and [44, 93]), the ADHD and the HC classes are not easily separable in either database. However, the heterogeneity of the population in these two databases, as well as the inconsistent acquisition protocols and the predefined test set in the ADHD-200 database, guarantee that a significant result in any of them, especially in the latter, would represent a step forward in the comprehension of the functional and effective connectivity causes behind ADHD.

In this study, macro-averaged f-measure instead of accuracy was optimized, avoiding the construction of MFC-like classifiers in most cases while allowing a good generalization to datasets where the balance of the classes was different i.e. from the manually defined ADHD-200 validation sets to its predefined test sets. The results shall be discussed next following the general order in which they were presented in Chapter 6. The reader is reminded that an evaluation of the ACPI results should be done with caution given the intrinsic positive bias of k-fold cross validation and especially of LOO cross-validation.

7.1 Individual methods

The works of Wang et al. [147] and Smith et al. [129], which compared several statistical methods to calculate brain connectivity, do not agree on which method is best for rs-fMRI data. To date, no method has been proven better than the others with such a consistency that one could safely opt for it in every situation. As suggested by those authors, the best method will likely depend on the case at hand, given that their response to noise depends heavily on the type of noise present in the data. Though, with increasingly better de-noising techniques, it remains to be seen if one method is going to start having an edge over the others. Given the non-linear nature of brain processes it seems unlikely, at first,

that a cascade of those processes results in a linear or quasi-linear relationship between two remote structures. However, linear methods such as Fourier based coherence and, mostly, correlation, have been widely used, often with success, to assess FC in, among others, EEG and fMRI signals. Also, stationarity has been assumed when determining FC with some of these methods, an assumption which has been rightfully questioned by some authors e.g. [83]. Here, all these subtleties have not been taken into consideration prior to classification since the main goal of this study was not to achieve the best possible accuracy or to find the most meaningful connections between regions for ADHD-HC discrimination but to see if extracting information from several of these methods would help this cause.

Results in the ACPI database clearly indicated that mutual information based methods were the ones that provided connectivity reconstructions with better discriminative power to SVM classifiers. The accuracy value of 74.4% achieved in said dataset represents a proof of concept example of the distinguishability of the two classes. Even if all other methods performed as badly as a random classifier, the results achieved with mutual information in this dataset would prove that the two classes are, in fact, distinguishable to a certain point through ML techniques. It is, however, not straightforward why mutual information achieved such results since in any of the two referred studies [129, 147] this method performed particularly well. However, it is not easy to attribute these results to the sensibility of mutual information to non linear relationships, since other methods with this characteristic, such as the BH2U, did not perform well in the same conditions. Given such results, one would look for a confirmation of the importance of this method for FC estimation, or, at least, for classification of ADHD subjects, in the results achieved in the ADHD-200 database. Unfortunately, this was not the case in the filtered test set which was the one of most interest. In the face of this inconsistency, the question now is if one should reject the hypotheses that mutual information plays an important role in distinguishing subjects with ADHD from HC. In classifications such as the ones performed in this study, the number of variables that take part in the final result is of such an order that a poor result can be a consequence of any of those variables and not necessarily due to the indistinguishability of the two classes. This reasoning is backed up by the results achieved with this method in every set but the filtered test set. In the validation sets of the filtered ADHD-200 dataset, in the non filtered validation and test sets and also, to some degree, in the classifications with RFs, mutual information based methods were the best or close to being it. In Section 6.1 a most probable overfitting hypothesis was already advanced as an explanation to this. The possibility that filtering might be prejudicial to classifications based on FC calculated with mutual information seems unlikely, since most information of FC has been suggested, on a strong basis, to be located in the frequency range 0.009-0.08 Hz [19, 34], though it should not be discarded. Further attempts at classifying the filtered test set with mutual information should be made to investigate if there were any possible irregularities in the process that would have caused the achieved results. Nonetheless, strong evidence of the value of mutual

information for classification of subjects with ADHD can be derived from the results of this study.

Since correlation is a very common method to evaluate FC we shall now elaborate on the achieved results with it. Correlation measures linear relationships between variables, essentially taking into account their variance. The correlation implemented in MULAN is not the Pearson's correlation coefficient, since it considers different time lags between the variables. Instead, the correlation implemented in MULAN is the maximum covariance value between the two centred and standardised variables at different lags, τ, $\tau < L$, where L is the maximum lag in time points, to be defined by the user. L is the only parameter that has to be defined for this method. In this study, $L = 12$, which means around 24 s in case a TR of 2 s was considered. This period of time seems enough to include any delayed relationship between two regions and, in fact, the maximum value of coherence occurred for $\tau = 0$ in most cases. Thus, not much more could have been done to improve the reconstructed connections with correlation apart from changing variables in the overall pipeline. In both Wang's [147] and Smith's [129] studies, correlation methods performed quite well, something confirmed by their overall performance in this study, actually originating the best results in some datasets, as in the ADHD-200 filtered test set without selection of features. As such, there is no evidence in this study that would not suggest the usage of correlation based methods to measure FC.

As already mentioned in Chapter 3, in Smith's study, one of the best methods was partial correlation. The partial version of the used methods calculated by matrix inversion, were not included in this study to avoid increasing the number of classifications to be made and to avoid using methods without having first tested their simpler versions. Partial versions of methods are one way to determine the statistical dependency between two random variables while controlling for the others, something already mentioned in Chapter 2. It would be interesting to see what impact the usage of the methods' partial versions instead of the bivariate ones would have on classification.

7.2 Impact of Feature Selection

When using a single matrix to extract features without selection, the classifiers had to deal with 13456 features. When five matrices were combined, that number was five times greater, meaning that the classifiers had to deal with a staggering 67280 features in such cases. Some classifiers handle high-dimensional feature spaces better than others, and so SVMs and RFs were chosen here because they were expected to work better than average in such feature sets. For RFs this was thought to be true because they use only a subset of features to construct each node, thus, making their final classification reliant on only a few features from the original set.

As already said, the problem with having a big number of features and a small sample is the resulting sparsity of data in such high dimensional feature sets. As the number

of features increases, the space between data points gets wider and the number of solutions to the classification problem increases exponentially [103]. Though, the large margin principle in which the SVM algorithm is built upon, results in meaningful decision boundaries even in high dimensional spaces. Besides that, the SVM algorithm defines a number of features from the relationships between the data points, meaning that, if the original feature set is discarded, the SVM algorithm is going to act just on this new, possibly smaller feature set, of constant size. Because of these two reasons SVMs also handle high-dimensional feature sets well.

Feature selection can certainly help reducing the number of solutions to the classification problem to a more meaningful set as discussed in Section 2.3.2, but how can it particularly help an SVM classifier? The trade-off between margin size and the number of margin violations allowed by the soft margin approach of SVMs to classification has to be balanced in such a way that the noise-related outliers are allowed to violate the margin, while keeping the largest margin between the meaningful data. This trade-off is defined by the regularization parameter C and possibly by some kernel specific parameters, which depend on distance-like measures. Because the behaviour of distance measures in high-dimensional spaces is hard to predict, reducing the dimensionality of the data can benefit SVMs by making it easier to find the optimal parameter values.

Having what was said into account the hypothesis would be that feature selection could reduce overfitting and possibly improve SVM classification performance.

Due to the large number of classifications to make and to the large number of features involved in them, the chosen feature selection approach had to be fast and able to drastically reduce the dimensionality of the feature space, while keeping the most meaningful features. As previously mentioned, a two step process was adopted in order to achieve this. A univariate feature selection was used to eliminate most of the features, followed by a further dimensionality reduction based on PCA. The overall results were mixed. In the ACPI dataset, results were, on average, slightly worse using feature selection, however, in the ADHD-200 dataset they were reasonably better, as hypothesised. Though, the best results in the ACPI dataset and in both ADHD-200 test sets were always achieved with feature selection. To explain the achieved results in the ACPI dataset one might need to look at how the used univariate feature selection works. The statistical tests used to select the features test the null hypotheses that the means of both classes' distributions are the same. Which p-value would give us a reasonable certainty that the two classes actually have a different mean? If all features came from the same distribution and in all of them the two classes had the same mean, in more than 67000 features almost certainly some would have lower p-values than what usually is considered a significant value (around 0.05). It is possible, thus, that a univariate feature selection of this type on such a great number of features would yield a subset in which some features would have a low p-value by chance alone. One hypothesis to explain the results in the ACPI dataset would be that in this dataset, the overall p-values of the 500 chosen features were higher than in the ADHD-200 dataset, which was, in fact, the case (average of ≈ 0.003 in the ACPI's best 500

features and ≈ 0.001 in the ones from the ADHD-200 filtered dataset). If the p-values were not sufficiently low, it could have been that the means of the two classes were only different in the training set but that this difference only happened by chance, meaning that it did not translate to new subjects and, thus, to the test set where the classifier was evaluated. Thus, only the two principal components of the data's PCA representation were used for classification in order to avoid overfitting the training set.

A last note on feature selection shall be made, specifically on why so many times the SVM classifiers acted as a MFC classifier in both ADHD-200 test sets, when feature selection was made. One expects a MFC-like classification if the classifier is in total underfit, however, in both cases in which the respective results in the validation set are reported, the macro-averaged f-measure was reasonably good, meaning that there was no underfit in the validation set. Even though this theory could not be tested during the period this study was made, this seemingly paradoxal behaviour could be explained as a consequence of the two linear transformations applied on the test sets. These two linear transformations were derived first from the scaling of the training set data and second from the PCA representation of the training set. If the training set data of the ADHD-200 dataset is not representative of the data present in the predefined test set, both linear transformations could project the test set data to a completely different part of the feature space than the one where the decision boundary was drawn, thus leading to all data points being in a single side of it and a MFC-like classification.

7.3 Impact of Filtering

The kind of filtering made in the ADHD-200 filtered dataset is common practice in rs-fMRI preprocessing. With a TR of 2 s the Nyquist–Shannon theorem states that any signal limited at 0.25 Hz can be totally reconstructed. Filtering the data in the frequency band 0.009-0.08/0.1 Hz removes respiratory artefacts (0.1-0.5 Hz [34]) included in the frequencies represented by the BOLD signal as well as other types of noise, leaving the part of the signal with most FC information. One would expect a confirmation of these previous findings [19, 34] in the classification results achieved in this study. As a matter of fact, that was the overall conclusion in Section 6.3.

In the validation sets results, filtering did not have a big impact on classification which can be a result of the overestimation that comes with optimizing the parameters in said data. Also, because part of the predefined test set comes from data acquired in different sites than that of the training set, noise not present in the training set data could be present in the test set data, and, therefore, if not filtered out, it could mislead the classifier to worse classification performances.

Even though the overall results in ADHD-200 test sets were considerably higher in the filtered test set, three of the four best results occurred in the filtered test set when feature selection was used. The reason as to why this happened and what role feature selection played in it remains to be investigated.

7.4 Combination of Matrices

As discussed in Chapter 2, every statistical method used to reconstruct the FC connectivity of an individual measures different aspects of the statistical dependency between each ROI's time series. The first idea behind combining the information of the connectivity matrices constructed with these methods was to take advantage of the different information each method contains about the true connectivity underlying the BOLD signal. In reality, however, ML goes beyond that. Ideally, a classifier should be able not only to use the best out of each method but also to find relationships between the different measured statistical dependencies. For example, it could find how the connectivity between regions A and B measured with a given method α relates with the connectivity between regions C and D measured with method β and tell us how that relationship relates to the class of the subject in analysis. Besides that, an ML algorithm could establish comparisons between methods if it allows us to recover the information about which features from each method it used to classify a given dataset. In this study, mostly because it represents one of the first steps towards understanding how a combination of matrices could help in classification, a simple but comprehensive approach of feeding the information directly from each matrix as a feature to the classifier was adopted. Though, other studies of pattern recognition could be made to extract information from this central idea. For instance, in this study, each feature was fed to the classifier as if the relationship between themselves was equal. That, however, is not true. The first weights of all matrices measure the connectivity between the same regions, as well as all other weights with the same vertical and horizontal indexes. This information could be given to the classifier or used for feature selection. Also, it would be interesting to understand what benefits the classifier more, a combination of similar methods or a combination of methods that measure different characteristics of signals.

The success of combining matrices for classification relies on the ability of the classifier to handle very high-dimensional feature spaces and/or on the quality of feature selection. Both of these problems are also posed for classification with a single matrix but are drastically emphasised when four or five matrices are used simultaneously. Most of the weights of each matrix were not of great use for discriminating both the ADHD and the HC class, as confirmed by the results achieved with feature selection using only one matrix, which were generally better even with a reduction of 13456 to only 80 features in the ADHD-200 dataset. If one supposes that of all constructed matrices the most discriminative set of features is composed of some features from each matrix, then, if the chosen feature selection consistently filters those out of the original set, one would expect, in this case, an overall increase in classification performance as the number of combined matrices increases. If, on the other hand, feature selection is not made prior to classification, then, due to the curse of dimensionality and the larger number of "noisy" features, the classification is expected to decrease.

In the ACPI dataset, when several matrices were combined, the average of the macro-averaged f-measure was lower and the best results were worse than when only a single matrix was used. This agrees with what was expected. When feature selection was made, the distributions were quite similar but the best results, achieved with mutual information, were not replicated when the matrix built with the best method of this family in the validation set was combined with other matrices. Why would this be? If feature selection was perfect, this selection would, at least, keep the features on which the classification with the mutual information method was based, maintaining the performance achieved using just the mutual information method. There are several reasons as to why this might not have happened. The major one is, probably, the kind of feature selection used in this study. If the results achieved with mutual information were a consequence of the relationship between various features, or of even just two, and those features were not sufficiently significant on their own in the training set to pass the ANOVA test, then, at least one of them, or all of them in the worst case scenario, would not be in the selected feature set in which PCA representation was determined. Surely, in this case, the classification performance would drop if features from other matrices were not included in that feature set. As the number of combined matrices increases, the number of features to be filtered in feature selection increases too, making it harder for the features responsible for the good performance of the mutual information method to be in the final group of 500 features selected with the significance test. The representativeness of the training set also has a big impact on whether the features needed to distinguish the test data are selected or not.

In the ADHD-200 test sets, when no feature selection was made, results were similar when using a single matrix and when using a combination of matrices, as mentioned in Chapter 6. Even the best results in either test set followed the same pattern, the best macro-averaged f-measure corresponded to a classification with a single matrix and the best accuracy value to one with a combination of matrices. When feature selection was made, the best results were much better than the best ones achieved without feature selection, and the combination of matrices clearly benefited classification in some cases. The drop in performance of the best methods from the non-selected case to the selected case can be explained by the same phenomenon caused by the univariate selection mentioned in the last paragraph. The BCohF[1] method seemed to help classification in almost all cases in the ADHD-200 filtered test set. Though, these results do not seem to be a result of the relationship between the weights of this matrix and the weights of the other matrices but rather of the good discriminative power of its most significant features on the test set. One would say this because all matrices paired well with the BCohF[1] matrix, leading one to think that the reason for the achieved results is the common denominator in all classifications. The alternative to this explanation, that there was an equally good relationship between the most significant features of the BCohF[1] matrix and the ones from each of the other matrices, seems unlikely. The best combination of matrices across all datasets of both databases was the combination of a mutual information method with

a transfer entropy method.

Even though some evidence that the combination of matrices with feature selection could help in the classification of subjects with ADHD, the four graphs in Figure 6.6 clearly show that when five matrices were combined the results were consistently poor, being often close to the worst performance achieved with a single matrix or being actually worse than that.

Taking into account the results achieved in this study and in the study made by Deshpande et al. [41], the combination of matrices is recommended when the goal is to achieve the best possible classification performances in a given dataset, while using feature selection. The use of single matrices should, nonetheless, keep being adopted at the same time to not compromise any hypothetical findings and to provide further comparisons between the two approaches. Potential downsides of using a combination of FC matrices for classification are mostly associated with finding the tools to apply the necessary statistical methods and the additional time spent on their application, on feature selection and on classification.

The type of feature selection used in this study, far from being the ideal one, was fast and allowed a significant reduction in the parameter optimisation and in the classifier training time. While permitting this, it was often able to collect a group of features that surpassed the equivalent classification performance without it. Though, after looking at the results achieved with the combination of matrices, one is led to think that such a big univariate selection of features might have compromised any beneficial relationship between the features of different matrices and also between the features of the same matrix. Thus, a smaller study to analyse how the type of feature selection affects the performance of a classifier when several matrices are combined is here proposed.

An ensemble-like classification in which several SVM estimators as the ones mentioned in this study would classify the database using a different matrix, voting between each other to decide the final classification is also proposed for further investigation. Because the number of estimators would probably be small, instead of the voting system, the class attribution probability each SVM classifier gives could be averaged to decide what the combined classification should be or, instead, the most extreme probability could be considered.

7.5 General Considerations

Even though, as already mentioned, the major goal of this work was not to achieve the best possible classification performances, the overall results reported here reveal how difficult it is to distinguish the ADHD and the HC classes. While DL methods promise to eliminate part of the feature engineering steps in the present pipeline of connectivity based classifications, it remains to be explored what is causing the less than optimal results in ADHD vs HC classifications of studies adopting it, particularly in the ADHD-200

dataset. Since there are so many inter-subject variables in the ADHD-200 database (acquisition site, resting-state instructions, age, gender, type of ADHD, session duration, TR, among others) it is not easy to identify what is causing this phenomenon. As some authors are already doing, de-constructing the database to assemble more homogeneous datasets might help this cause. In the author's point of view, the consequences of averaging possibly very heterogeneous time series to form a single time series per atlas' ROI should be investigated, while comparing them with more homogeneous ROIs defined by taking into account the data in hand. Also, based on the results of Wang's study [147], longer rs-fMRI sessions could greatly benefit the accuracy of the statistical methods used to reconstruct each subject's brain connectivity. In addition, it remains to be studied how classification performances of each method relate to their ability to reconstruct the true functional and effective connectivity of each subject, as only some aspects of the relationship between different brain regions might be interesting to distinguish some classes of subjects.

In this study, only SVM classifiers were tested with feature selection because RFs have a type of feature selection of their own. However, this does not mean that RFs would not benefit from prior feature selection because the random feature subset from which each node has to be chosen would possibly be more meaningful in that case. Nonetheless, the results of this study do not support the use of RFs over SVMs in the type of datasets considered here.

Finally, due to time limitations, two further analysis were left to be made. First, statistical tests to compare the results between the several variations introduced in this study would have been a valuable complement to the analysis presented here. A paired t-test or a Wilcoxon signed-rank test would have been possible choices for this matter. Second, it was not possible to analyse which connections between the different AAL defined regions of the brain were most often used to distinguish both classes, which had more weight in the decision function of the SVM classifier, or which were consistently given higher importance values by the RF classifiers, though, those kinds of analysis are of great importance to connect the ML world and the biomedicine world and should be done whenever possible.

Chapter 8

Conclusions

In this work, evidence was provided that the combination of FC matrices can be used to improve the performance of ADHD vs HC classifications, especially when prior feature selection is made. Nonetheless, no particular combination of matrices was found to significantly outperform the others on a consistent basis. Additionally, mutual information based methods were shown to play an important role in such classifications, namely when using data from the ACPI and the ADHD-200 public databases. Also, the achieved results support the importance of the BOLD signal in the frequency range 0.009-0.08 Hz for brain connectivity based classifications. ANOVA testing followed by PCA dimensionality reduction was proven to be an efficient feature selection technique in the classifications performed with SVMs, as well. This type of feature selection, however, might be too reliant on univariate feature selection to be able to maximize classification performance. Finally, RF classifiers performed significantly worse than SVMs. As such, their use over the latter is not suggested in the conditions the classifications in this work were made.

The classification performances previously reported in the ACPI database [93] were surpassed. A maximum accuracy and macro-averaged f-measure of 0.744 and 0.677 were achieved, respectively. Also, the best accuracy results of 0.61 achieved in the ADHD-200 database competition held in 2011 [4] were also surpassed [44], using the publicly available test set for classifier evaluation. A maximum accuracy and macro-averaged f-measure of 0.678 and 0.648 were achieved in this case, respectively. In spite of these results being good relative to the current literature, they are still far from what a classifier should achieve in order to be helpful in clinical practice.

Given the conclusions presented here, more research on the combination of FC matrices should be made to assess the extent to which they can further improve classification performances on subjects with ADHD or other psychiatric, and neurological diseases and to explore the effect of other feature selection techniques besides the one used in this

study in such cases. In addition, it remains to be confirmed the value of mutual information for classification of subjects with ADHD on databases besides the ACPI and the ADHD-200.

Bibliography

[1] *ACPI*. 2015. URL: http://fcon_1000.projects.nitrc.org/indi/ACPI/html/index.html (visited on 07/09/2017).

[2] *ACPI scan parameters*. URL: https://s3.amazonaws.com/fcp-indi/data/Projects/ACPI/ScanParameters/mta_1_scan_parameters.pdf (visited on 09/07/2017).

[3] P. S. Addison. *The Illustrated Wavelet Transform Handbook: Introdutory Theory and Applications in Science, Engineering, Medicine and Finance*. 2nd Edition. Boca Raton: Taylor & Francis Group, 2017, p. 446.

[4] *ADHD-200*. URL: http://fcon_1000.projects.nitrc.org/indi/adhd200/index.html (visited on 07/08/2017).

[5] A. M.H. J. Aertsen and G. L. Gerstein. "Evaluation of neuronal connectivity: Sensitivity of cross-correlation." In: *Brain Research* 340.2 (1985), pp. 341–354.

[6] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. "On the Surprising Behavior of Distance Metrics in High Dimensional Spaces." In: *Proceedings of the 8th International Conference on Database Theory*. ICDT '01. London, UK, UK: Springer-Verlag, 2001, pp. 420–434.

[7] Analysis Group. *FSL website*. 2017. URL: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/.

[8] J. S. Anderson, J. A. Nielsen, A. L. Froehlich, M. B. DuBray, T. J. Druzgal, A. N. Cariello, J. R. Cooperrider, B. A. Zielinski, C. Ravichandran, P. T. Fletcher, A. L. Alexander, E. D. Bigler, N. Lange, and J. E. Lainhart. "Functional connectivity magnetic resonance imaging classification of autism." In: *Brain* 134.12 (2011), pp. 3742–3754.

[9] M. R. Arbabshirani, E. Castro, and V. D. Calhoun. "Accurate classification of schizophrenia patients based on novel resting-state fMRI features." In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2014, pp. 6691–6694.

[10] A. P. Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[11] *Athena Pipeline webpage*. 2011. URL: http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline.

[12] S. Aydin, N. Arica, E. Ergul, and O. Tan. "Classification of Obsessive Compulsive Disorder by EEG Complexity and Hemispheric Dependency Measurements." In: *International Journal of Neural Systems* 25.03 (2015), 1550010(1)–1550010(16).

[13] P. A. Bandettini. "Twenty years of functional MRI: The science and the stories." In: *NeuroImage* 62.2 (2012), pp. 575–588.

[14] A. M. Bastos and J.-m. Schoffelen. "A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls." In: *Frontiers in systems neuroscience* 9.January (2016), pp. 1–23.

[15] P. Bellec, C. Chu, F. Chouinard-decorte, Y. Benhajali, D. S. Margulies, and R. C. Craddock. "NeuroImage The Neuro Bureau ADHD-200 Preprocessed repository." In: *NeuroImage* 144 (2017), pp. 275–286.

[16] R. Bellman. *Adaptive Control Processes: A Guided Tour*. 1st Edition. Princeton University Press, 1961, p. 94.

[17] J. S. Bendat and A. G. Persol. *Random Data*. Ed. by D. J. Balding, N. A. C. Cressie, G. M. Fitzmaurice, I. M. Johnstone, G. Molenberghs, D. W. Scott, A. F. M. Smith, R. S. Tsay, and W. S. 4th Edition. New Jersey: John Wiley & Sons, 2010, p. 604.

[18] C. M. Bishop. *Pattern Recognition and Machine Learning*. Ed. by M. Jordan, J. Kleinbeg, and B. Schölkopf. 1st Edition. Springer, 2006, p. 738.

[19] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde. "Functional Connectivity in the Motor Cortex of Resting Human Brain Using Echo-Planar MRI." In: *Magnetic resonance in medicine* 34.4 (1995), pp. 537–41.

[20] K. J. Blinowska. "Review of the methods of determination of directed connectivity from multichannel data." In: *Medical & Biological Engineering & Computing* 49 (2011), pp. 521–529.

[21] L. Breiman. "Bagging predictors." In: *Machine Learning* 24.2 (1996), pp. 123–140.

[22] L. Breiman. "Random Forests." In: *Machine Learning* 45.1 (2001), pp. 5–32.

[23] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone. *Classification and regression trees*. 1st Edition. hapman and Hall/CRC, 1984.

[24] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan. *MRI Physical Principles and Sequence Design*. 2nd Edition. New Jersey: John Wiley & Sons, 2014, p. 944.

[25] C. J. C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition." In: *Data Mining and Knowledge Discovery* 2.2 (1998), pp. 121–167.

[26] *C-PAC*. URL: http://fcp-indi.github.io/ (visited on 07/09/2017).

[27] M. S. Cetin, J. M. Houck, V. M. Vergara, R. L. Miller, and V. Calhoun. "Multimodal based classification of schizophrenia patients." In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2015, pp. 2629–2632.

[28] *Child Mind Institute - INDI*. 2017. URL: http://fcon_1000.projects.nitrc.org/ (visited on 07/09/2017).

[29] C. Chu, A. L. Hsu, K. H. Chou, P. Bandettini, and C. P. Lin. "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images." In: *NeuroImage* 60.1 (2012), pp. 59–70.

[30] S. Chu, L. Ekström, and R. Firestone. *The Lund/LBNL Nuclear Data Search*. 1999. URL: http://nucleardata.nuclear.lu.se/toi/ (visited on 10/06/2017).

[31] Y.-s. Chung, D. F. Hsu, C.-Y. Liu, and C.-Y. Tang. "Performance evaluation of classifier ensembles in terms of diversity and performance of individual systems." In: *International Journal of Pervasive Computing and Communications* 6.4 (2010), pp. 373–403.

[32] D. Chyzhyk, M. Graña, D. Öngür, and A. K. Shinn. "Discrimination of Schizophrenia Auditory Hallucinators by Machine Learning of Resting-State Functional MRI." In: *International Journal of Neural Systems* 25.03 (2015), 1550007(1)–1550007(23).

[33] A. L. Cohen, D. A. Fair, N. U. F. Dosenbach, F. M. Miezin, D. Dierker, D. C. V. Essen, B. L. Schlaggar, and S. E. Petersen. "Defining functional areas in individual human brains using resting functional connectivity MRI." In: *NeuroImage* 41.1 (2008), pp. 45–57.

[34] D. Cordes, V. M. Haughton, K. Arfanakis, J. D. Carew, P. A. Turski, C. H. Moritz, M. A. Quigley, and M. E. Meyerand. "Frequencies contributing to functional connectivity in the cerebral cortex in "resting-state"data." In: *American Journal of Neuroradiology* 22.7 (2001), pp. 1326–1333.

[35] C. Cortes and V. Vapnik. "Support-vector networks." In: *Machine Learning* 20.3 (1995), pp. 273–297.

[36] R. C. Craddock, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg. "Disease state prediction from resting state functional connectivity." In: *Magnetic Resonance in Medicine* 62.6 (2009), pp. 1619–1628.

[37] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg. "A whole brain fMRI atlas generated via spatially constrained spectral clustering." In: *Human Brain Mapping* 33.8 (2012), pp. 1914–1928.

[38] O. David, D. Cosmelli, and K. J. Friston. "Evaluation of different measures of functional connectivity using a neural mass model." In: *NeuroImage* 21.2 (2004), pp. 659–673.

[39] G. Deco and M. L. Kringelbach. "Great expectations: Using whole-brain computational connectomics for understanding neuropsychiatric disorders." In: *Neuron* 84.5 (2014), pp. 892–905.

[40] G. Deshpande, K. Sathian, and X. Hu. "Assessing and Compensating for Zero-Lag Correlation Effects in Time-Lagged Granger Causality Analysis of fMRI." In: *IEEE Transactions on Biomedical Engineering* 57.6 (2010), pp. 1446–1456.

[41] G. Deshpande, P. Wang, D. Rangaprakash, and B. Wilamowski. "Fully Connected Cascade Artificial Neural Network Architecture for Attention Deficit Hyperactivity Disorder Classification From Functional Magnetic Resonance Imaging Data." In: *IEEE Transactions on Cybernetics* 45.12 (2015), pp. 2668–2679.

[42] S. Dey, A. R. Rao, and M. Shah. "Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects." In: *Frontiers in Neural Circuits* 8.June (2014), pp. 1–11.

[43] T. Elomaa. "The Biases of Decision Tree Pruning Strategies." In: *Advances in Intelligent Data Analysis: Third International Symposium, IDA-99 Amsterdam, The Netherlands, August 9–11, 1999 Proceedings*. Ed. by D. J. Hand, J. N. Kok, and M. R. Berthold. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 63–74.

[44] A. Eloyan, J. Muschelli, M. B. Nebel, H. Liu, F. Han, T. Zhao, A. D. Barber, S. Joel, J. J. Pekar, S. H. Mostofsky, and B. Caffo. "Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging." In: *Frontiers in Systems Neuroscience* 6.August (2012), pp. 1–9.

[45] A. D. Elster. *Questions and Answers in MRI*. 2017. URL: http://mri-q.com/rapid-imaging-fse-epi.html (visited on 10/06/2017).

[46] D. C. V. Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, K. Ugurbil, and W.-m.H.C. P. Consortium. "NeuroImage The WU-Minn Human Connectome Project : An overview." In: *NeuroImage* 80 (2013), pp. 62–79.

[47] A. A. Fingelkurts, A. A. Fingelkurts, and S. Kähkönen. "Functional connectivity in the brain - is it an elusive concept ?" In: *Neuroscience and Biobehavioral Reviews* 28 (2005), pp. 827–836.

[48] M. D. Fox and M. E. Raichle. "Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging." In: *Nat Rev Neurosci* 8.9 (2007), pp. 700–711.

[49] M. D. Fox and M. Greicius. "Clinical applications of resting state functional connectivity." In: *Front Syst Neurosci* 4.19 (2010), pp. 1–13.

[50] Y. Freund and R. E. Schapire. "A desicion-theoretic generalization of on-line learning and an application to boosting." In: *Computational Learning Theory: Second European Conference, EuroCOLT '95 Barcelona, Spain, March 13–15, 1995 Proceedings*. Ed. by P. Vitányi. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 23–37.

[51] A. Frid. "Differences in phase synchrony of brain regions between regular and dyslexic readers." In: *2014 IEEE 28th Convention of Electrical Electronics Engineers in Israel (IEEEI)*. 2014, pp. 1–4.

[52] J. H. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine." In: *The Annals of Statistics* 29.5 (2001), pp. 1189–1232.

[53] K. J. Friston, L. Harrison, and W. Penny. "Dynamic causal modelling." In: *NeuroImage* 19.4 (2003), pp. 1273–1302.

[54] K. J. Friston. "Functional and Effective Connectivity : A Review." In: *Brain Connectivity* 1.1 (2011), pp. 13–36.

[55] R. Garcia, E. C. Paraiso, and J. C. Nievola. "Comparative Study of Dimensionality Reduction Methods Using Reliable Features for Multiple Datasets Obtained by rs-fMRI in ADHD Prediction." In: *Advances in Artificial Intelligence: 30th Canadian Conference on Artificial Intelligence, Canadian AI 2017, Edmonton, AB, Canada, May 16-19, 2017, Proceedings*. Ed. by M. Mouhoub and P. Langlais. Cham: Springer International Publishing, 2017, pp. 97–102.

[56] F. Ge, J. Lv, X. Hu, B. Ge, L. Guo, J. Han, and T. Liu. "Deriving ADHD biomarkers with sparse coding based network analysis." In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. 2015, pp. 22–25.

[57] A. Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. Ed. by N. Tache. 1st Edition. O'Reilly Media, 2017, p. 568.

[58] A. Gholipour, N. Kehtarnavaz, R. Briggs, M. Devous, and K. Gopinath. "Brain Functional Localization : A Survey of Image Registration Techniques." In: *IEEE Transactions on Medical Imaging* 26.4 (2007), pp. 427–451.

[59] J. D. Gispert, J. Pascau, S. Reig, R. Martínez-Lázaro, V. Molina, P. García-Barreno, and M. Desco. "Influence of the normalization template on the outcome of statistical parametric mapping of PET scans." In: *NeuroImage* 19.3 (2003), pp. 601–612.

[60] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. 1st Edition. The MIT press, 2016, p. 785.

[61] C. W. J. Granger. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods." In: *Econometrica* 37.3 (1969), pp. 424–438.

[62] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon. "Functional connectivity in the resting brain: a network analysis of the default mode hypothesis." In: *Proceedings of the National Academy of Sciences of the United States of America* 100.1 (2003), pp. 253–8.

[63] A. Grinsted, J. C. Moore, and S. Jevrejeva. "Application of the cross wavelet transform and wavelet coherence to geophysical time series." In: *Nonlinear Processes in Geophysics, European Geosciences Union (EGU)* 11.5/6 (2004), pp. 561–566.

[64] A. J. Hao, B. L. He, and C. H. Yin. "Discrimination of ADHD children based on Deep Bayesian Network." English. In: *2015 IET International Conference on Biomedical Image and Signal Processing (ICBISP 2015)*. Institution of Engineering and Technology, 2015, pp. 1–6.

[65] J. M. Harlow. "Recovery from the passage of an iron bar through the head." In: *History of Psychiatry* 4.14 (1993), pp. 274–281.

[66] H. Hart, K. Chantiluke, A. I. Cubillo, A. B. Smith, A. Simmons, M. J. Brammer, A. F. Marquand, and K. Rubia. "Pattern classification of response inhibition in ADHD: Toward the development of neurobiological markers for ADHD." In: *Human Brain Mapping* 35.7 (2014), pp. 3083–3094.

[67] H. Hart, A. F. Marquand, A. Smith, A. Cubillo, A. Simmons, M. Brammer, and K. Rubia. "Predictive Neurofunctional Markers of Attention-Deficit/Hyperactivity Disorder Based on Pattern Classification of Temporal Processing." In: *Journal of the American Academy of Child & Adolescent Psychiatry* 53.5 (2014), 569 –578.e1.

[68] D. Heath, S. Kasif, and S. Salzberg. "Induction of Oblique Decision Trees." In: *Journal of Artificial Intelligence Research* 2.2 (1993), pp. 1–32.

[69] M. P. van den Heuvel and H. E. Hulshoff Pol. "Exploring the brain network: A review on resting-state fMRI functional connectivity." In: *European Neuropsychopharmacology* 20.8 (2010), pp. 519–534.

[70] T. K. Ho. "Random decision forests." In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1. 1995, pp. 278–282.

[71] J. Hua, W. D. Tembe, and E. R. Dougherty. "Performance of feature-selection methods in the classification of high-dimension data." In: *Pattern Recognition* 42.3 (2009), pp. 409–424.

[72] S. A. Huettel, A. W. Song, and G. McCarthy. *Functional Magnetic Resonance Imaging*. 2nd Edition. Sunderland: Sinauer Associates Inc, 2009, p. 542.

[73] G. Hughes. "On the mean accuracy of statistical pattern recognizers." In: *IEEE Transactions on Information Theory* 14.1 (1968), pp. 55–63.

[74] L. Igual, J. C. Soliva, S. Escalera, R. Gimeno, O. Vilarroya, and P. Radeva. "Automatic brain caudate nuclei segmentation and classification in diagnostic of Attention-Deficit/Hyperactivity Disorder." In: *Computerized Medical Imaging and Graphics* 36.8 (2012), pp. 591–600.

[75] B. Jie, D. Zhang, W. Gao, Q. Wang, C. Y. Wee, and D. Shen. "Integration of Network Topological and Connectivity Properties for Neuroimaging Classification." In: *IEEE Transactions on Biomedical Engineering* 61.2 (2014), pp. 576–589.

[76] N. F. Jie, M. H. Zhu, X. Y. Ma, E. A. Osuch, M. Wammes, J. Theberge, H. D. Li, Y. Zhang, T. Z. Jiang, J. Sui, and V. D. Calhoun. "Discriminating Bipolar Disorder From Major Depression Based on SVM-FoBa: Efficient Feature Selection With Multimodal Brain Imaging Data." In: *IEEE Transactions on Autonomous Mental Development* 7.4 (2015), pp. 320–331.

[77] I. T. Jolliffe. *Principal Component Analysis*. 2nd Edition. Springer, 2002, p. 487.

[78] S. Kapur, A. G. Phillips, and T. R. Insel. "Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it?" In: *Molecular Psychiatry* 17.12 (2012), pp. 1174–1179.

[79] S. M. Kay and S. L. Marple. "Spectrum Analysis-A Modern Perspective." In: *Proceedings of the IEEE* 69.11 (1981), pp. 1380–1419.

[80] K. Krane. *Modern Physics*. Ed. by S. Johnson. 3rd Edition. John Wiley & Sons, 2012, p. 550.

[81] D. Kuang and L. He. "Classification on ADHD with deep learning." In: *Proceedings - 2014 International Conference on Cloud Computing and Big Data, CCBD 2014* (2014), pp. 27–32.

[82] S. Kullback. *Statistics and Information theory*. J. Wiley and Sons, New York, 1959.

[83] J.-p. Lachaux, A. Lutz, D. Rudrauf, D. Cosmelli, M. L. V. Quyen, J. Martinerie, and F. Varela. "Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence." In: *Clinical Neurophysiology* 32 (2002), pp. 157–174.

[84] L. D. Landau, E. M. Lifshitz, and L. P. Pitaevskii. *Course of Theoretical Physics Volume 9: Statistical Physics Part 2*. 3rd Edition. Oxford: Pergamon Press Inc., 1980, p. 387.

[85] D. T. Larose. "k-Nearest Neighbor Algorithm." In: *Discovering Knowledge in Data*. John Wiley & Sons, Inc., 2005, pp. 90–106.

[86] Y. Lecun, Y. Bengio, and G. Hinton. "Deep learning." In: *Nature* 521.7553 (2015), pp. 436–444.

[87] A. F. Leuchter, T. F. Newton, I. A. Cook, D. O. Walter, S. Rosenberg-Thompson, and P. A. Lachenbruch. "Changes in Brain Functional Connectivity in Alzheimer-type and Multi-infarct Dementia." In: *Brain* 115.5 (1992), pp. 1543–1561.

[88] Y. Li, Z. Lian, M. Li, Z. Liu, L. Xiao, and Z. Wei. "ELM-based classification of ADHD patients using a novel local feature extraction method." In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2016, pp. 489–492.

[89] F. Lopes da Silva, J. P. Pijn, and P. Boeijinga. "Interdependence of EEG signals: Linear vs. nonlinear associations and the significance of time delays and phase shifts." In: *Brain Topography* 2.1-2 (1989), pp. 9–18.

[90] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi. "A review of classification algorithms for EEG-based brain–computer interfaces." In: *Journal of Neural Engineering* 4.2 (2007), R1–R13.

[91] I. G. Maglogiannis. *Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies*. Vol. 160. Ios Press, 2007, pp. 12–13.

[92] D. Marinazzo, M. Pellicoro, and S. Stramaglia. "Kernel Method for Nonlinear Granger Causality." In: *Phys. Rev. Lett.* 100.14 (2008), 144103(1)–144103(4).

[93] R. Meszlényi, L. Peska, V. Gál, Z. Vidnyánszky, and K. Buza. "Classification of fMRI data using Dynamic Time Warping based functional connectivity analysis." In: *European Signal Processing Conference*. IEE, 2016, pp. 245–249.

[94] R. Moran, D. A. Pinotsis, and K. Friston. "Neural masses and fields in dynamic causal modeling." In: *Frontiers in Computational Neuroscience* 7 (2013), pp. 1–12.

[95] *MTA 1*. 2015. URL: http://fcon_1000.projects.nitrc.org/indi/ACPI/html/acpi_mta_1.html (visited on 07/09/2017).

[96] P. Mukherjee, J. I. Berman, S. W. Chung, C. P. Hess, and R. G. Henry. "Diffusion tensor MR imaging and fiber tractography: Theoretic underpinnings." In: *American Journal of Neuroradiology* 29.4 (2008), pp. 632–641.

[97] *MULAN*. 2016. URL: https://github.com/HuifangWang/MULAN (visited on 05/10/2017).

[98] A. C. Müller and S. Guido. *Introduction to Machine Learning with Python*. Ed. by D. Schanafelt. 1st Edition. O'Reilly Media, 2017, p. 394.

[99] K. P. Murphy. *Machine Learning A Probabilistic Perspective*. 1st Edition. The MIT press, 2012, p. 1067.

[100] K. Murphy, R. M. Birn, D. A. Handwerker, T. B. Jones, and P. A. Bandettini. "The impact of global signal regression on resting state correlations : Are anti-correlated networks introduced ?" In: *NeuroImage* 44.3 (2009), pp. 893–905.

[101] B. Mwangi, T. S. Tian, and J. C. Soares. "A Review of Feature Reduction Techniques in Neuroimaging." In: *Neuroinformatics* 12.2 (2014), pp. 229–244.

[102] J. Nielsen, B. Zielinski, P Fletcher, A. Alexander, N. Lange, E. Bigler, J. Lainhart, and J. Anderson. "Multisite functional connectivity MRI classification of autism: ABIDE results." In: *Frontiers in Human Neuroscience* 7 (2013), 599(1)–599(12).

[103] W. S. Noble. "What is a support vector machine?" In: *Nature Biotechnology* 24.12 (2006), pp. 1565–1567.

[104] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank. "Brain magnetic resonance imaging with contrast dependent on blood oxygenation." In: *Proceedings of the National Academy of Sciences of the United States of America* 87.24 (1990), pp. 9868–72.

[105] R. S. Patel, F. D. Bowman, and J. K. Rilling. "A Bayesian approach to determining connectivity of the human brain." In: *Human Brain Mapping* 27.3 (2006), pp. 267–276.

[106] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[107] X. Peng, P. Lin, T. Zhang, and J. Wang. "Extreme Learning Machine-Based Classification of ADHD Using Brain Structural MRI Data." In: *PLOS ONE* 8.11 (2013), pp. 1–12.

[108] D. H. Perkel, G. L. Gerstein, and G. P. Moore. "Neuronal Spike Trains and Stochastic Point Processes: II. Simultaneous Spike Trains." In: *Biophysical Journal* 7.4 (1967), pp. 419–440.

[109] L. Pollonini, U. Patidar, N. Situ, R. Rezaie, A. C. Papanicolaou, and G. Zouridakis. "Functional connectivity networks in the autistic and healthy brain assessed using Granger causality." In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. 2010, pp. 1730–1733.

[110] D. M. W. Powers. "Evaluation: From Precision, Recall and F-Measure To ROC, Informedness, Markedness & Correlation." In: *Journal of Machine Learning Technologies* 2.1 (2011), pp. 37–63.

[111] *Preprocessed Connectomes Project website*. URL: http://preprocessed-connectomes-project.org/ (visited on 09/07/2017).

[112] E. M. Purcell, H. C. Torrey, and R. V. Pound. "Resonance Absorption by Nuclear Magnetic Moments in a Solid." In: *Physical Review* 69.1-2 (1946), pp. 37–38.

[113] J. R. Quinlan. "Induction of decision trees." In: *Machine Learning* 1.1 (1986), pp. 81–106.

[114] J. R. Quinlan. *C4.5: Programs for Machine Learning*. 1st Edition. Morgan Kaufmann, 1993, p. 302.

[115] M. N. I. Qureshi, H. J. Jo, and B. Lee. "ADHD subgroup discrimination with global connectivity features using hierarchical extreme learning machine: Resting-state FMRI study." In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. 2017, pp. 529–532.

[116] D. Rangaprakash, X. Hu, and G. Deshpande. "Phase Synchronization in Brain Networks Derived From Correlation Between Probabilities of Recurrences in Functional Data." In: *International Journal of Neural Systems* 23.2 (2013), 1350003(1)–1350003(11).

[117] J. L. Rodgers and W. A. Nicewander. "Thirteen Ways to Look at the Correlation Coefficient." In: *The American Statistician* 42.1 (1988), pp. 59–66.

[118] M. S. Roulston. "Estimating the errors on measured entropy and mutual information." In: *Physica D* 125.5759 (1999), pp. 285–294.

[119] Y. Saeys, I. Inza, and P. Larrañaga. "A review of feature selection techniques in bioinformatics." In: *Bioinformatics* 23.19 (2007), pp. 2507–2517.

[120] V. Sakkalis. "Review of advanced techniques for the estimation of brain connectivity measured with EEG / MEG." In: *Computers in Biology and Medicine* 41.12 (2011), pp. 1110–1117.

[121] T. Schreiber. "Measuring Information Transfer." In: *Physical Review Letters* 85.2 (2000), pp. 461–464.

[122] Scikit-learn Developers. *Scikit-Learn documentation*. 2017. URL: http://scikit-learn.org/stable/documentation.

[123] B. Sen, G. A. Bernstein, T. Xu, B. A. Mueller, M. W. Schreiner, K. R. Cullen, and K. K. Parhi. "Classification of obsessive-compulsive disorder from resting-state fMRI." In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2016, pp. 3606–3609.

[124] C. E. Shannon and W. Weaver. *The Mathematical Theory of Information*. Urbana, Illinois: University of Illinois Press, 1949.

[125] S. K. Shenas, U. Halici, and M. Çiçek. "A comparative analysis of functional connectivity data in resting and task-related conditions of the brain for disease signature of OCD." In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2014, pp. 978–981.

[126] J. Shlens. "A Tutorial on Principal Component Analysis." In: *CoRR* abs/1404.1100 (2014), pp. 1–12.

[127] J. F. Smith, K. Chen, A. S. Pillai, and B. Horwitz. "Identifying effective connectivity parameters in simulated fMRI: A direct comparison of switching linear dynamic system, stochastic dynamic causal, and multivariate autoregressive models." In: *Frontiers in Neuroscience* 7.MAY (2013), pp. 1–17.

[128] N. B. Smith and A. Webb. *Introduction to Medical Imaging: Physics, Engineering and Clinical Applications*. 1st Edition. New York: Cambridge University Press, 2011, p. 286.

[129] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. "Network modelling methods for FMRI." In: *NeuroImage* 54.2 (2011), pp. 875–891.

[130] S. Song, Z. Zhan, Z. Long, J. Zhang, and L. Yao. "Comparative Study of SVM Methods Combined with Voxel Selection for Object Category Classification on fMRI Data." In: *PLOS ONE* 6.2 (2011), pp. 1–11.

[131] O. Sporns. "Brain connectivity." In: *Scholarpedia* 2.10 (2007), p. 4695.

[132] O. Sporns. "Connectome." In: *Scholarpedia* 5.2 (2010), p. 5584.

[133] O. Sporns, G. Tononi, and R. Kötter. "The human connectome: A structural description of the human brain." In: *PLoS Computational Biology* 1.4 (2005), pp. 0245–0251.

[134] J. Stelzer, Y. Chen, and R. Turner. "Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control." In: *NeuroImage* 65 (2013), pp. 69–82.

[135] J. Tang, C. Deng, and G.-B. Guang. "Extreme learning machine for multilayer perceptron." In: *IEEE Transactions on Neural Networks and Learning Systems* 27.4 (2015), pp. 809–821.

[136] The INDI Team. *ACPI preprocessing pipelines*. 2015. URL: http://fcon_1000.projects.nitrc.org/indi/ACPI/html/preproc.html (visited on 09/07/2017).

[137] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288.

[138] R. Tibshirani. "Regression shrinkage and selection via the lasso: a retrospective." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3 (2011), pp. 273–282.

[139] C. Torrence and G. P. Compo. "A Practical Guide to Wavelet Analysis." In: *Bulletin of the American Meteorological Society* 79.1 (1998), pp. 61–78.

[140] G. Tsoumakas, I. Katakis, and I. Vlahavas. "Mining Multi-label Data." In: *Data Mining and Knowledge Discovery Handbook*. Ed. by O. Maimon and L. Rokach. Boston, MA: Springer US, 2010, pp. 667–685.

[141] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain." In: *NeuroImage* 15.1 (2002), pp. 273–289.

[142] L. J. P. Van Der Maaten and G. E. Hinton. "Visualizing high-dimensional data using t-sne." In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[143] J. D. Van Horn and A. W. Toga. "Brain Atlases: Their Development and Role in Functional Inference." In: *fMRI Techniques and Protocols*. Ed. by M. Filippi. New York, NY: Springer New York, 2016, pp. 265–281.

[144] J. A. Vastano and H. L. Swinney. "Information transport in spatiotemporal systems." In: *Phys. Rev. Lett.* 60 (18 1988), pp. 1773–1776.

[145] R. Vicente, M. Wibral, M. Lindner, and G. Pipa. "Transfer entropy-a model-free measure of effective connectivity for the neurosciences." In: *Journal of Computational Neuroscience* 30.1 (2011), pp. 45–67.

[146] S. Vieira, W. H. Pinaya, and A. Mechelli. "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications." In: *Neuroscience and Biobehavioral Reviews* 74 (2017), pp. 58–75.

[147] H. E. Wang, C. G. Bénar, P. Pascale, K. J. Friston, V. K. Jirsa, P. A. Valdes-sosa, and J. W. Bohland. "A systematic framework for functional connectivity measures." In: *Frontiers in Neuroscience* 8.December (2014), pp. 1–22.

[148] X. Wang, Y. Jiao, and Z. Lu. "Discriminative analysis of resting-state brain functional connectivity patterns of Attention-Deficit Hyperactivity Disorder using Kernel Principal Component Analysis." In: *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. Vol. 3. 2011, pp. 1938–1941.

[149] X. Wang, Y. Jiao, T. Tang, H. Wang, and Z. Lu. "Altered regional homogeneity patterns in adults with attention-deficit hyperactivity disorder." In: *European Journal of Radiology* 82.9 (2013), pp. 1552–1557.

[150] Y. Wang, J. Ji, and P. Liang. "Feature selection of fMRI data based on normalized mutual information and fisher discriminant ratio." In: *Journal of X-Ray Science and Technology* 24.3 (2016), pp. 467–475.

[151] M. Wibral, R. Vicente, and J. T. Lizier. *Directed Information Measures in Neuroscience*. Ed. by M. Wibral, R. Vicente, and J. T. Lizier. 1st Edition. Heidelberg: Springer, 2014, p. 225.

[152] N. Wiener. "The theory of prediction." In: *Modern mathematics for engineers* 1 (1956), pp. 125–139.

[153] T. Wolfers, J. K. Buitelaar, C. F. Beckmann, B. Franke, and A. F. Marquand. "From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics." In: *Neuroscience and Biobehavioral Reviews* 57 (2015), pp. 328–349.

[154] Y. Zang, T. Jiang, Y. Lu, Y. He, and L. Tian. "Regional homogeneity approach to fMRI data analysis." In: *NeuroImage* 22.1 (2004), pp. 394–400.

[155] L. L. Zeng, H. Shen, L. Liu, L. Wang, B. Li, P. Fang, Z. Zhou, Y. Li, and D. Hu. "Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis." In: *Brain* 135.5 (2012), pp. 1498–1507.

[156] L. L. Zeng, H. Shen, L. Liu, and D. Hu. "Unsupervised classification of major depression using functional connectivity MRI." In: *Human Brain Mapping* 35.4 (2014), pp. 1630–1641.

[157] X. Zhang, B. Hu, X. Ma, and L. Xu. "Resting-State Whole-Brain Functional Connectivity Networks for MCI Classification Using L2-Regularized Logistic Regression." In: *IEEE Transactions on NanoBioscience* 14.2 (2015), pp. 237–247.

[158] C. Z. Zhu, Y. F. Zang, Q. J. Cao, C. G. Yan, Y. He, T. Z. Jiang, M. Q. Sui, and Y. F. Wang. "Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder." In: *NeuroImage* 40.1 (2008), pp. 110–120.

[159] H. Zou and T. Hastie. "Regularization and variable selection via the elastic net." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

[160] Q. Zou, J. Zeng, L. Cao, and R. Ji. "A novel features ranking metric with application to scalable visual and bioinformatics data classification." In: *Neurocomputing* 173 (2016), pp. 346–354.

[161] M. H. Zweig and G. Campbell. "Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine." In: *Clinical Chemistry* 39.4 (1993), pp. 561–577.

ANNEX I

Results - Complementary Tables

In this Annex the results of every type of classification performed in this study are presented. When a type of classification was performed more than once the average results and the corresponding standard deviation are presented. Best accuracy and macro-averaged f-measure values are in bold. Codes in tables are as follows: BCorrU, undirected bivariate correlation; BCorrD, directed bivariate correlation; BCohF[0], bivariate Fourier based coherence in frequency 0.01 Hz; BCohF[1], bivariate Fourier based coherence in frequency 0.043 Hz; BCohF[2], bivariate Fourier based coherence in frequency 0.076 Hz; BCohW, bivariate wavelet based coherence; BH2U, undirected bivariate h^2; BH2D, directed bivariate h^2; BMITU, undirected bivariate mutual information; BMITD1, directed bivariate mutual information; BMITD2, directed bivariate mutual information (corrected); BTEU, undirected bivariate transfer entropy; BTED, directed bivariate transfer entropy; acc, accuracy; fm, macro-averaged f-measure; TPR, true positive rate; prec, precision; std, standard deviation; est, number of estimators of best classification; C, value of C parameter; RP_fm, macro-averaged f-measure value of an equivalent classification using a coin-toss like classifier; WRP_acc, accuracy value of an equivalent classification using a classifier that assigns random classes in the proportion the classes exist in the evaluation set; MFC, accuracy value of an equivalent classification using a MFC classifier.

I.1 Results in the ACPI dataset

Table I.1: 5-fold CV with an RF classifier

Methods	est	acc	std	fm	std	TPR	std	prec	std
BCorrU	5	0.610	0.036	0.504	0.046	0.786	0.046	0.686	0.023
BCorrD	7	**0.629**	0.038	0.504	0.055	0.827	0.047	0.689	0.024
BCohF[0]	3	0.586	0.039	0.499	0.043	0.736	0.046	0.681	0.024
BCohF[1]	3	0.584	0.034	0.498	0.037	0.731	0.052	0.681	0.021
BCohF[2]	5	0.607	0.038	0.498	0.043	0.787	0.052	0.683	0.022
BCohW	3	0.589	0.039	0.500	0.042	0.741	0.054	0.682	0.024
BH2U	3	0.598	0.036	**0.509**	0.049	0.750	0.042	0.688	0.026
BH2D	5	0.602	0.040	0.495	0.045	0.779	0.050	0.681	0.024
BMITU	3	0.589	0.038	0.501	0.042	0.740	0.055	0.683	0.024
BMITD1	3	0.589	0.033	0.505	0.040	0.733	0.043	0.685	0.023
BMITD2	5	0.600	0.040	0.494	0.044	0.775	0.056	0.681	0.023
BTEU	3	0.588	0.047	0.499	0.054	0.740	0.049	0.681	0.031
BTED	5	0.612	0.034	0.502	0.043	0.794	0.039	0.685	0.021
BCorrD_BCohW	3	0.597	0.038	0.508	0.048	0.751	0.040	0.687	0.026
BCorrD_BH2U	5	0.617	0.037	0.508	0.044	0.799	0.045	0.688	0.022
BCorrD_BMITD1	5	0.611	0.032	0.508	0.038	0.784	0.044	0.688	0.019
BCorrD_BTED	3	0.585	0.041	0.496	0.039	0.739	0.058	0.679	0.022
BCohW_BH2U	5	0.608	0.037	0.500	0.041	0.788	0.052	0.684	0.020
BCohW_BMITD1	3	0.586	0.041	0.497	0.048	0.738	0.050	0.681	0.028
BCohW_BTED	3	0.588	0.038	0.499	0.047	0.740	0.043	0.682	0.026
BH2U_BMITD1	5	0.609	0.040	0.504	0.049	0.786	0.046	0.686	0.024
BH2U_BTED	5	0.616	0.042	0.505	0.048	0.801	0.051	0.687	0.024
BMITD1_BTED	3	0.587	0.035	0.502	0.043	0.732	0.046	0.683	0.025
BCorrD_BCohW_BH2U	3	0.589	0.038	0.500	0.044	0.740	0.051	0.683	0.026
BCorrD_BCohW_BMITD1	3	0.592	0.037	0.507	0.039	0.740	0.049	0.685	0.022
BCorrD_BCohW_BTED	3	0.595	0.032	0.505	0.036	0.749	0.047	0.685	0.020
BCorrD_BH2U_BMITD1	3	0.593	0.038	0.503	0.042	0.748	0.046	0.684	0.023
BCorrD_BH2U_BTED	3	0.590	0.041	0.501	0.045	0.742	0.052	0.682	0.025
BCorrD_BMITD1_BTED	3	0.589	0.044	0.497	0.046	0.745	0.056	0.680	0.025
BCohW_BH2U_BMITD1	5	0.608	0.030	0.499	0.041	0.788	0.037	0.684	0.021
BCohW_BH2U_BTED	3	0.589	0.035	0.493	0.043	0.749	0.049	0.679	0.024
BCohW_BMITD1_BTED	3	0.582	0.032	0.491	0.036	0.737	0.044	0.677	0.021
BH2U_BMITD1_BTED	3	0.588	0.046	0.497	0.050	0.743	0.059	0.680	0.028
BCorrD_BCohW_BH2U_BMITD1	3	0.590	0.038	0.502	0.043	0.743	0.048	0.683	0.023
BCorrD_BCohW_BH2U_BTED	3	0.587	0.039	0.497	0.038	0.740	0.061	0.680	0.022
BCorrD_BCohW_BMITD1_BTED	5	0.604	0.042	0.494	0.051	0.784	0.054	0.681	0.025
BCorrD_BH2U_BMITD1_BTED	7	0.621	0.036	0.500	0.046	0.816	0.045	0.686	0.021
BCohW_BH2U_BMITD1_BTED	3	0.586	0.042	0.500	0.049	0.734	0.051	0.682	0.028
BCorrD_BCohW_BH2U_BMITD1_BTED	5	0.607	0.035	0.495	0.045	0.789	0.044	0.682	0.022

RP_fm=0.482 std=0.045, WRP_acc=0.565 std=0.040, MFC=0.68

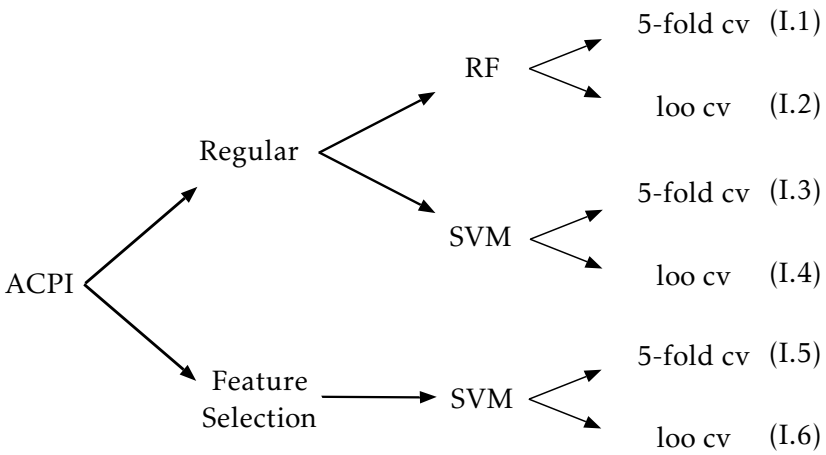


Table I.2: LOO CV with an RF classifier

Methods	est	acc	std	fm	std	TPR	std	prec	std
BCorrU	3	0.594	0.040	0.506	0.046	0.745	0.041	0.685	0.026
BCorrD	3	0.592	0.047	0.506	0.051	0.740	0.057	0.685	0.029
BCohF[0]	3	0.600	0.044	**0.513**	0.051	0.751	0.048	0.689	0.029
BCohF[1]	3	0.585	0.043	0.498	0.047	0.735	0.053	0.680	0.027
BCohF[2]	3	0.581	0.042	0.495	0.046	0.728	0.046	0.678	0.026
BCohW	3	0.587	0.037	0.497	0.042	0.741	0.042	0.680	0.024
BH2U	5	0.608	0.042	0.497	0.047	0.791	0.050	0.682	0.024
BH2D	3	0.584	0.036	0.498	0.038	0.731	0.052	0.681	0.022
BMITU	3	0.582	0.043	0.495	0.049	0.733	0.047	0.679	0.028
BMITD1	7	0.609	0.037	0.491	0.046	0.800	0.045	0.681	0.022
BMITD2	5	0.607	0.039	0.495	0.047	0.791	0.042	0.682	0.023
BTEU	3	0.587	0.038	0.493	0.043	0.747	0.052	0.679	0.023
BTED	5	0.599	0.033	0.489	0.039	0.780	0.041	0.678	0.020
BCorrD_BCohW	3	0.584	0.034	0.496	0.040	0.736	0.040	0.679	0.022
BCorrD_BH2U	3	0.586	0.046	0.497	0.053	0.739	0.049	0.680	0.030
BCorrD_BMITD1	3	0.582	0.044	0.493	0.054	0.734	0.048	0.679	0.030
BCorrD_BTED	3	0.589	0.043	0.501	0.043	0.740	0.055	0.682	0.025
BCohW_BH2U	3	0.590	0.038	0.497	0.046	0.749	0.045	0.681	0.025
BCohW_BMITD1	3	0.583	0.050	0.496	0.054	0.733	0.055	0.679	0.031
BCohW_BTED	5	0.601	0.041	0.492	0.049	0.781	0.048	0.680	0.025
BH2U_BMITD1	3	0.581	0.042	0.493	0.048	0.731	0.047	0.678	0.028
BH2U_BTED	3	0.587	0.041	0.495	0.049	0.744	0.042	0.679	0.027
BMITD1_BTED	3	0.584	0.040	0.500	0.046	0.731	0.041	0.681	0.027
BCorrD_BCohW_BH2U	3	0.592	0.037	0.502	0.048	0.747	0.039	0.684	0.026
BCorrD_BCohW_BMITD1	3	0.588	0.041	0.499	0.045	0.742	0.049	0.681	0.025
BCorrD_BCohW_BTED	3	0.585	0.031	0.492	0.034	0.744	0.044	0.678	0.019
BCorrD_BH2U_BMITD1	3	0.585	0.036	0.496	0.045	0.739	0.044	0.680	0.025
BCorrD_BH2U_BTED	5	**0.613**	0.032	0.500	0.038	0.799	0.039	0.685	0.019
BCorrD_BMITD1_BTED	5	0.600	0.038	0.490	0.045	0.781	0.048	0.679	0.023
BCohW_BH2U_BMITD1	3	0.582	0.041	0.493	0.047	0.736	0.049	0.678	0.026
BCohW_BH2U_BTED	5	0.599	0.038	0.489	0.040	0.781	0.050	0.678	0.021
BCohW_BMITD1_BTED	3	0.580	0.036	0.495	0.036	0.727	0.050	0.678	0.021
BH2U_BMITD1_BTED	3	0.590	0.039	0.497	0.043	0.748	0.046	0.680	0.024
BCorrD_BCohW_BH2U_BMITD1	5	0.597	0.031	0.487	0.043	0.779	0.031	0.678	0.021
BCorrD_BCohW_BH2U_BTED	3	0.586	0.035	0.495	0.042	0.742	0.046	0.679	0.023
BCorrD_BCohW_BMITD1_BTED	5	0.608	0.039	0.498	0.043	0.791	0.051	0.683	0.022
BCorrD_BH2U_BMITD1_BTED	5	0.612	0.040	0.499	0.044	0.799	0.047	0.683	0.022
BCohW_BH2U_BMITD1_BTED	3	0.585	0.049	0.501	0.049	0.733	0.058	0.681	0.030
BCorrD_BCohW_BH2U_BMITD1_BTED	7	0.622	0.037	0.495	0.048	0.825	0.040	0.684	0.022

RP_fm=0.482 std=0.045, WRP_acc=0.565 std=0.040, MFC=0.68

Table I.3: 5-fold CV with an SVM classifier and without feature selection

Methods	C	acc	std	fm	std	TPR	std	prec	std
BCorrU	15	0.647	0.026	0.474	0.032	0.896	0.031	0.683	0.013
BCorrD	19	0.647	0.025	0.491	0.025	0.883	0.035	0.687	0.011
BCohF[0]	22.5	0.680	0.021	0.535	0.030	0.911	0.024	0.705	0.012
BCohF[1]	65.5	0.619	0.024	0.466	0.030	0.848	0.030	0.675	0.012
BCohF[2]	62	0.610	0.020	0.430	0.025	0.864	0.022	0.665	0.010
BCohW	100	0.612	0.023	0.447	0.029	0.850	0.029	0.668	0.012
BH2U	122	0.641	0.020	0.481	0.023	0.878	0.026	0.683	0.010
BH2D	16	0.640	0.026	0.539	0.034	0.815	0.025	0.703	0.017
BMITU	2732	**0.705**	0.026	**0.632**	0.033	0.844	0.026	0.752	0.018
BMITD1	1912	0.694	0.026	0.606	0.037	0.856	0.027	0.737	0.020
BMITD2	1700	0.684	0.019	0.557	0.027	0.896	0.026	0.713	0.011
BTEU	17	0.669	0.015	0.451	0.023	0.959	0.016	0.684	0.007
BTED	22	0.671	0.017	0.451	0.024	0.962	0.022	0.685	0.007
BCorrD_BCohF[0]	23	0.673	0.021	0.509	0.031	0.919	0.025	0.697	0.011
BH2D_BMITU	100	0.640	0.027	0.554	0.032	0.793	0.030	0.711	0.017
BCohF[0]_BTED	70	0.672	0.013	0.476	0.025	0.944	0.019	0.689	0.007
BMITU_BTED	47	0.668	0.019	0.456	0.026	0.952	0.024	0.685	0.008
BCorrD_BMITU	30	0.652	0.026	0.494	0.031	0.891	0.031	0.689	0.012
BCohF[0]_BMITU	89	0.663	0.028	0.534	0.035	0.873	0.031	0.703	0.016
BCohF[0]_BH2D	62	0.632	0.021	0.504	0.028	0.838	0.029	0.689	0.012
BCorrD_BH2D	16	0.637	0.023	0.496	0.030	0.857	0.029	0.687	0.012
BH2D_BTED	16	0.638	0.024	0.508	0.035	0.845	0.024	0.691	0.015
BCorrD_BTED	65.5	0.667	0.017	0.466	0.027	0.942	0.019	0.686	0.009
BCorrD_BCohF[0]_BH2D	16	0.635	0.026	0.483	0.033	0.865	0.032	0.683	0.014
BCorrD_BCohF[0]_BMITU	26	0.665	0.022	0.498	0.030	0.913	0.028	0.693	0.011
BCorrD_BCohF[0]_BTED	19.5	0.671	0.019	0.468	0.026	0.946	0.021	0.687	0.009
BCorrD_BMITU_BTED	24.5	0.661	0.015	0.458	0.023	0.936	0.017	0.683	0.008
BH2D_BMITU_BTED	47	0.625	0.019	0.499	0.026	0.828	0.025	0.686	0.011
BCohF[0]_BH2D_BTED	18	0.639	0.024	0.488	0.028	0.870	0.028	0.685	0.012
BCohF[0]_BH2D_BMITU	30	0.630	0.027	0.508	0.032	0.828	0.029	0.689	0.015
BCorrD_BH2D_BMITU	19.5	0.631	0.024	0.492	0.031	0.848	0.027	0.684	0.013
BCohF[0]_BMITU_BTED	38	0.669	0.020	0.476	0.034	0.938	0.022	0.689	0.011
BCorrD_BH2D_BTED	17	0.635	0.023	0.473	0.028	0.874	0.028	0.680	0.012
BCorrD_BCohF[0]_BH2D_BMITU	22	0.631	0.026	0.477	0.031	0.862	0.033	0.680	0.013
BCorrD_BCohF[0]_BH2D_BTED	14	0.651	0.019	0.479	0.025	0.901	0.025	0.685	0.009
BCorrD_BCohF[0]_BMITU_BTED	26	0.672	0.013	0.477	0.027	0.942	0.014	0.689	0.008
BCorrD_BH2D_BMITU_BTED	23	0.631	0.021	0.470	0.025	0.869	0.026	0.679	0.010
BCohF[0]_BH2D_BMITU_BTED	25	0.634	0.020	0.484	0.024	0.863	0.029	0.683	0.010
BCorrD_BCohF[0]_BH2D_BMITU_BTED	20	0.638	0.019	0.466	0.027	0.887	0.022	0.679	0.010
								RP_fm=0.482 std=0.045, WRP_acc=0.565 std=0.040, MFC=0.68	

Table I.4: LOO CV with an SVM classifier and without feature selection

Methods	C	acc	fm	TPR	prec
BCorrU	15	0.632	0.484	0.859	0.682
BCorrD	19	0.632	0.484	0.859	0.682
BCohF[0]	22.5	0.688	0.545	0.918	0.709
BCohF[1]	65.5	0.624	0.452	0.871	0.673
BCohF[2]	62	0.616	0.433	0.871	0.667
BCohW	70	0.640	0.429	0.918	0.672
BH2U	122	0.632	0.471	0.871	0.679
BH2D	16	0.632	0.535	0.800	0.701
BMITU	2732	**0.728**	**0.663**	0.859	0.768
BMITD1	1912	0.720	0.650	0.859	0.760
BMITD2	1700	0.656	0.523	0.871	0.698
BTEU	17	0.672	0.463	0.953	0.686
BTED	22	0.664	0.459	0.941	0.684
BCorrD_BCohF[0]	23	0.640	0.489	0.871	0.685
BH2D_BMITU	100	0.640	0.565	0.776	0.717
BCohF[0]_BTED	70	0.656	0.470	0.918	0.684
BMITU_BTED	47	0.672	0.480	0.941	0.690
BCorrD_BMITU	30	0.600	0.439	0.835	0.664
BCohF[0]_BMITU	89	0.664	0.539	0.871	0.705
BCohF[0]_BH2D	62	0.592	0.447	0.812	0.663
BCorrD_BH2D	16	0.624	0.479	0.847	0.679
BH2D_BTED	16	0.624	0.501	0.824	0.686
BCorrD_BTED	65.5	0.680	0.500	0.941	0.696
BCorrD_BCohF[0]_BH2D	16	0.600	0.464	0.812	0.670
BCorrD_BCohF[0]_BMITU	26	0.640	0.489	0.871	0.685
BCorrD_BCohF[0]_BTED	19.5	0.664	0.459	0.941	0.684
BCorrD_BMITU_BTED	24.5	0.672	0.480	0.941	0.690
BH2D_BMITU_BTED	47	0.624	0.501	0.824	0.686
BCohF[0]_BH2D_BTED	18	0.624	0.466	0.859	0.676
BCohF[0]_BH2D_BMITU	30	0.600	0.464	0.812	0.670
BCorrD_BH2D_BMITU	19.5	0.632	0.507	0.835	0.689
BCohF[0]_BMITU_BTED	38	0.664	0.475	0.929	0.687
BCorrD_BH2D_BTED	17	0.624	0.479	0.847	0.679
BCorrD_BCohF[0]_BH2D_BMITU	22	0.600	0.464	0.812	0.670
BCorrD_BCohF[0]_BH2D_BTED	14	0.632	0.457	0.882	0.676
BCorrD_BCohF[0]_BMITU_BTED	26	0.648	0.450	0.918	0.678
BCorrD_BH2D_BMITU_BTED	23	0.624	0.479	0.847	0.679
BCohF[0]_BH2D_BMITU_BTED	25	0.608	0.456	0.835	0.670
BCorrD_BCohF[0]_BH2D_BMITU_BTED	20	0.632	0.457	0.882	0.676
RP_fm=0.482 std=0.045, WRP_acc=0.565 std=0.040, MFC=0.68					

Table I.5: 5-fold CV with an SVM classifier and with feature selection

Methods	C	acc	std	fm	std	TPR	std	prec	std
BCorrU	6	0.621	0.036	0.479	0.041	0.833	0.075	0.680	0.016
BCorrD	12	0.613	0.046	0.495	0.054	0.804	0.064	0.683	0.026
BCohF[0]	30	0.606	0.044	0.503	0.051	0.779	0.052	0.685	0.026
BCohF[1]	0.925	0.637	0.028	0.460	0.041	0.888	0.033	0.678	0.015
BCohF[2]	1.4	0.619	0.034	0.486	0.049	0.827	0.042	0.681	0.021
BCohW	26.8	0.609	0.041	0.501	0.049	0.787	0.052	0.685	0.025
BH2U	175	0.589	0.049	0.495	0.055	0.748	0.062	0.680	0.030
BH2D	2	0.602	0.041	0.494	0.047	0.780	0.057	0.681	0.024
BMITU	3.3	0.601	0.036	0.487	0.044	0.786	0.044	0.678	0.022
BMITD1	1	0.641	0.031	0.479	0.042	0.881	0.038	0.683	0.016
BMITD2	2.7	0.601	0.037	0.495	0.048	0.777	0.036	0.681	0.024
BTEU	1.22	0.634	0.030	0.478	0.046	0.865	0.043	0.682	0.018
BTED	1.5	0.642	0.030	0.472	0.046	0.886	0.053	0.683	0.016
BCorrD_BCohF[0]	4.7	0.602	0.027	0.482	0.031	0.794	0.042	0.676	0.015
BCorrD_BH2U	8.1	0.608	0.035	0.501	0.041	0.787	0.045	0.684	0.021
BCorrD_BMITD2	1.9	0.624	0.033	0.503	0.040	0.820	0.039	0.687	0.019
BCorrD_BTEU	1.85	0.618	0.030	0.485	0.041	0.825	0.044	0.681	0.017
BCohF[0]_BH2U	6.7	0.584	0.031	0.470	0.040	0.770	0.042	0.669	0.019
BCohF[0]_BMITD2	2.1	0.602	0.033	0.482	0.041	0.795	0.039	0.676	0.019
BCohF[0]_BTEU	7.6	0.617	0.035	0.505	0.045	0.803	0.035	0.687	0.022
BH2U_BMITD2	29.7	0.575	0.032	0.479	0.033	0.737	0.044	0.670	0.019
BH2U_BTEU	10.4	0.612	0.036	0.500	0.045	0.796	0.042	0.684	0.022
BMITD2_BTEU	1.2	**0.651**	0.026	0.485	0.041	0.895	0.038	0.687	0.015
BCorrD_BCohF[0]_BH2U	2.6	0.602	0.022	0.479	0.024	0.799	0.034	0.675	0.012
BCorrD_BCohF[0]_BMITD2	7	0.589	0.029	0.473	0.034	0.776	0.037	0.671	0.017
BCorrD_BCohF[0]_BTEU	6.7	0.617	0.031	0.495	0.043	0.813	0.038	0.684	0.020
BCorrD_BH2U_BMITD2	2.4	0.616	0.032	0.493	0.043	0.813	0.037	0.683	0.020
BCorrD_BH2U_BTEU	1.8	0.627	0.026	0.492	0.040	0.839	0.034	0.684	0.016
BCorrD_BMITD2_BTEU	2	0.627	0.032	0.492	0.045	0.841	0.035	0.684	0.019
BCohF[0]_BH2U_BMITD2	1.5	0.598	0.027	0.469	0.032	0.801	0.039	0.671	0.015
BCohF[0]_BH2U_BTEU	45	0.611	0.025	0.495	0.035	0.799	0.034	0.683	0.017
BCohF[0]_BMITD2_BTEU	3.2	0.624	0.031	**0.512**	0.036	0.811	0.041	0.690	0.017
BH2U_BMITD2_BTEU	3.1	0.610	0.029	0.479	0.037	0.815	0.038	0.677	0.017
BCorrD_BCohF[0]_BH2U_BMITD2	12	0.585	0.035	0.472	0.040	0.769	0.042	0.669	0.021
BCorrD_BCohF[0]_BH2U_BTEU	25.9	0.611	0.032	0.495	0.042	0.800	0.039	0.682	0.020
BCorrD_BCohF[0]_BMITD2_BTEU	9	0.608	0.032	0.488	0.043	0.800	0.042	0.680	0.020
BCorrD_BH2U_BMITD2_BTEU	12	0.623	0.031	0.512	0.038	0.807	0.038	0.691	0.018
BCohF[0]_BH2U_BMITD2_BTEU	9	0.614	0.029	0.493	0.037	0.808	0.039	0.682	0.018
BCorrD_BCohF[0]_BH2U_BMITD2_BTEU	28	0.609	0.035	0.492	0.044	0.799	0.044	0.681	0.021
				RP_fm=0.482 std=0.045, WRP_acc=0.565 std=0.040, MFC=0.68					

Table I.6: LOO CV with an SVM classifier and with feature selection

Methods	C	acc	fm	TPR	prec
BCorrU	6	0.552	0.471	0.694	0.663
BCorrD	12	0.632	0.551	0.776	0.71
BCohF[0]	30	0.624	0.538	0.776	0.702
BCohF[1]	0.925	0.584	0.402	0.835	0.651
BCohF[2]	1.4	0.6	0.452	0.824	0.667
BCohW	26.8	0.528	0.446	0.671	0.648
BH2U	175	0.56	0.468	0.718	0.663
BH2D	2	0.608	0.51	0.776	0.688
BMITU	3.3	0.64	0.557	0.788	0.713
BMITD1	1	0.616	0.515	0.788	0.691
BMITD2	2.7	**0.744**	**0.677**	0.882	0.773
BTEU	1.22	0.696	0.592	0.882	0.728
BTED	1.5	0.632	0.425	0.906	0.67
BCorrD_BCohF[0]	4.7	0.568	0.464	0.741	0.663
BCorrD_BH2U	8.1	0.624	0.552	0.753	0.711
BCorrD_BMITD2	1.9	0.632	0.507	0.835	0.689
BCorrD_BTEU	1.85	0.592	0.447	0.812	0.663
BCohF[0]_BH2U	6.7	0.576	0.47	0.753	0.667
BCohF[0]_BMITD2	2.1	0.624	0.53	0.788	0.698
BCohF[0]_BTEU	7.6	0.64	0.541	0.812	0.704
BH2U_BMITD2	29.7	0.592	0.49	0.765	0.677
BH2U_BTEU	10.4	0.592	0.498	0.753	0.681
BMITD2_BTEU	1.2	0.672	0.545	0.882	0.708
BCorrD_BCohF[0]_BH2U	2.6	0.6	0.439	0.835	0.664
BCorrD_BCohF[0]_BMITD2	7	0.592	0.498	0.753	0.681
BCorrD_BCohF[0]_BTEU	6.7	0.592	0.447	0.812	0.663
BCorrD_BH2U_BMITD2	2.4	0.632	0.507	0.835	0.689
BCorrD_BH2U_BTEU	1.8	0.584	0.442	0.8	0.66
BCorrD_BMITD2_BTEU	2	0.576	0.438	0.788	0.657
BCohF[0]_BH2U_BMITD2	1.5	0.608	0.491	0.8	0.68
BCohF[0]_BH2U_BTEU	45	0.6	0.464	0.812	0.67
BCohF[0]_BMITD2_BTEU	3.2	0.624	0.511	0.812	0.69
BH2U_BMITD2_BTEU	3.1	0.568	0.455	0.753	0.66
BCorrD_BCohF[0]_BH2U_BMITD2	12	0.592	0.421	0.835	0.657
BCorrD_BCohF[0]_BH2U_BTEU	25.9	0.536	0.452	0.682	0.652
BCorrD_BCohF[0]_BMITD2_BTEU	9	0.6	0.495	0.776	0.68
BCorrD_BH2U_BMITD2_BTEU	12	0.632	0.544	0.788	0.705
BCohF[0]_BH2U_BMITD2_BTEU	9	0.584	0.442	0.8	0.66
BCorrD_BCohF[0]_BH2U_BMITD2_BTEU	28	0.536	0.443	0.694	0.648
RP_fm=0.482 std=0.045, WRP_acc=0.565 std=0.040, MFC=0.68					

I.2 Results in the ADHD-200 datasets

Table I.7: RF classifier in the validation set

Methods	est	acc	std	fm	std	TPR	std	prec	std
BCorrU	3	0.572	0.041	0.517	0.044	0.336	0.071	0.381	0.064
BCorrD	5	0.586	0.040	0.515	0.044	0.292	0.070	0.388	0.072
BCohF[0]	3	0.559	0.039	0.498	0.043	0.301	0.071	0.355	0.064
BCohF[1]	5	0.572	0.037	0.496	0.041	0.263	0.066	0.357	0.068
BCohF[2]	3	0.564	0.041	0.504	0.045	0.311	0.073	0.363	0.068
BCohW	5	0.582	0.039	0.514	0.043	0.298	0.069	0.383	0.068
BH2U	3	0.571	0.040	0.515	0.043	0.331	0.073	0.379	0.063
BH2D	5	0.584	0.037	0.512	0.042	0.286	0.066	0.382	0.069
BMITU	5	0.609	0.037	0.545	0.042	0.338	0.073	0.432	0.067
BMITD1	7	**0.614**	0.037	0.544	0.043	0.319	0.071	0.437	0.071
BMITD2	3	0.602	0.039	**0.548**	0.042	0.368	0.070	0.429	0.064
BTEU	5	0.585	0.040	0.512	0.045	0.284	0.068	0.384	0.074
BTED	3	0.571	0.042	0.511	0.045	0.319	0.070	0.375	0.067

RP_fm=0.488 std=0.043, WRP_acc=0.543 std=0.042, MFC=0.647

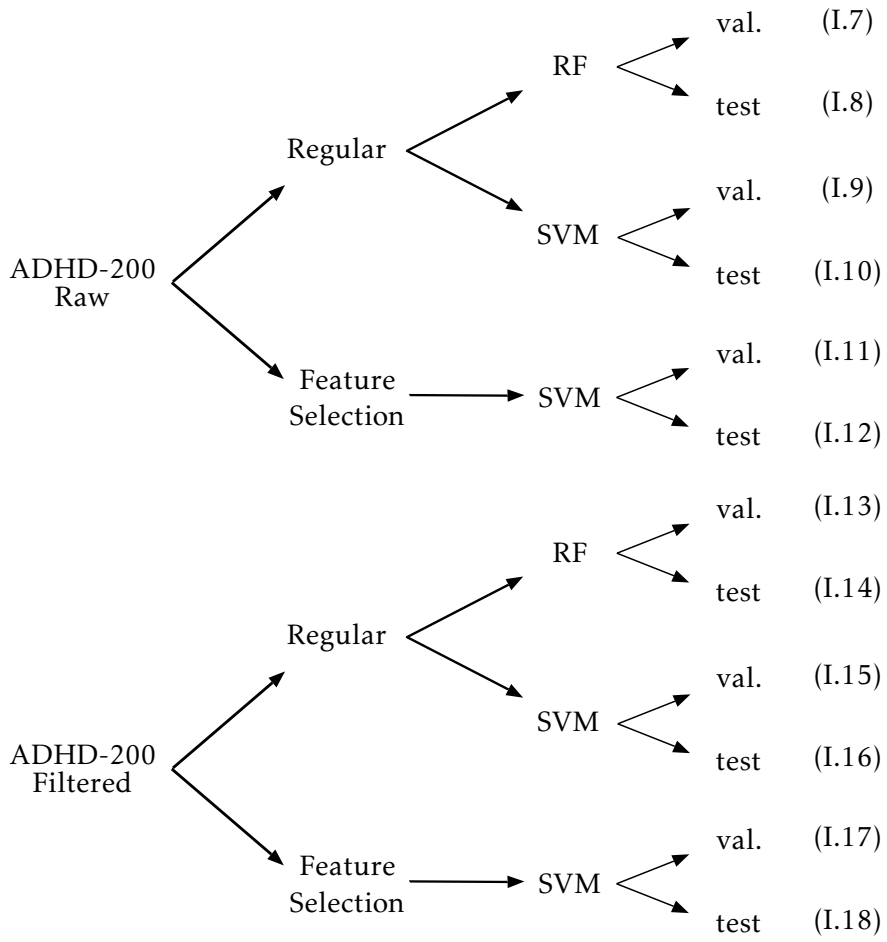


Table I.8: RF classifier in the test set

Methods	est	acc	std	fm	std	TPR	std	prec	std
BCorrU	3	0.501	0.041	0.473	0.042	0.306	0.062	0.429	0.063
BCorrD	3	0.523	0.041	0.492	0.044	0.312	0.061	0.459	0.066
BCohF[0]	3	0.515	0.035	0.481	0.037	0.293	0.054	0.445	0.057
BCohF[1]	3	0.521	0.036	0.488	0.039	0.302	0.057	0.455	0.061
BCohF[2]	3	0.509	0.036	0.476	0.039	0.293	0.061	0.435	0.058
BCohW	3	0.515	0.040	0.488	0.041	0.317	0.053	0.450	0.065
BH2U	3	0.501	0.042	0.475	0.043	0.313	0.060	0.430	0.063
BH2D	3	0.493	0.038	0.466	0.041	0.301	0.061	0.416	0.061
BMITU	3	0.521	0.034	0.491	0.036	0.313	0.054	0.456	0.055
BMITD1	5	0.530	0.034	0.492	0.039	0.288	0.065	0.466	0.062
BMITD2	5	**0.534**	0.037	**0.495**	0.041	0.286	0.056	0.474	0.068
BTEU	3	0.512	0.040	0.484	0.043	0.312	0.064	0.443	0.064
BTED	3	0.507	0.041	0.479	0.043	0.309	0.061	0.436	0.064
BCorrD_BCohW	3	0.519	0.037	0.490	0.040	0.315	0.058	0.453	0.061
BCorrD_BH2U	3	0.513	0.043	0.486	0.046	0.318	0.065	0.446	0.068
BCorrD_BMITD2	3	0.516	0.036	0.483	0.041	0.300	0.066	0.445	0.062
BCorrD_BTEU	3	0.516	0.034	0.487	0.037	0.313	0.061	0.449	0.055
BCohW_BH2U	5	0.522	0.034	0.488	0.037	0.300	0.057	0.456	0.057
BCohW_BMITD2	3	0.517	0.036	0.487	0.039	0.310	0.063	0.449	0.061
BCohW_BTEU	3	0.513	0.042	0.489	0.045	0.332	0.067	0.448	0.064
BH2U_BMITD2	3	0.511	0.036	0.484	0.037	0.317	0.059	0.444	0.054
BH2U_BTEU	3	0.503	0.041	0.475	0.042	0.307	0.058	0.431	0.063
BMITD2_BTEU	3	0.521	0.038	0.490	0.040	0.307	0.058	0.457	0.064
BCorrD_BCohW_BH2U	3	0.513	0.041	0.486	0.042	0.320	0.055	0.447	0.062
BCorrD_BCohW_BMITD2	3	0.517	0.038	0.488	0.039	0.313	0.055	0.452	0.059
BCorrD_BCohW_BTEU	3	0.519	0.044	0.493	0.048	0.330	0.067	0.455	0.069
BCorrD_BH2U_BMITD2	3	0.517	0.041	0.484	0.041	0.299	0.056	0.450	0.067
BCorrD_BH2U_BTEU	3	0.507	0.041	0.482	0.041	0.319	0.057	0.440	0.061
BCorrD_BMITD2_BTEU	3	0.517	0.041	0.485	0.043	0.302	0.063	0.449	0.065
BCohW_BH2U_BMITD2	3	0.512	0.044	0.483	0.045	0.307	0.055	0.445	0.068
BCohW_BH2U_BTEU	3	0.513	0.035	0.488	0.036	0.326	0.051	0.447	0.051
BCohW_BMITD2_BTEU	3	0.515	0.040	0.488	0.042	0.322	0.059	0.449	0.063
BH2U_BMITD2_BTEU	3	0.512	0.036	0.483	0.038	0.308	0.059	0.442	0.056
BCorrD_BCohW_BH2U_BMITD2	3	0.519	0.039	0.488	0.042	0.306	0.061	0.451	0.063
BCorrD_BCohW_BH2U_BTEU	3	0.514	0.032	0.487	0.033	0.322	0.059	0.448	0.049
BCorrD_BCohW_BMITD2_BTEU	5	0.526	0.038	0.486	0.041	0.278	0.062	0.460	0.069
BCorrD_BH2U_BMITD2_BTEU	3	0.515	0.037	0.486	0.039	0.311	0.059	0.448	0.059
BCohW_BH2U_BMITD2_BTEU	3	0.509	0.042	0.483	0.044	0.317	0.063	0.441	0.063
BCorrD_BCohW_BH2U_BMITD2_BTEU	3	0.512	0.040	0.483	0.043	0.313	0.065	0.443	0.066

RP_fm=0.498 std=0.043, WRP_acc=0.504 std=0.042, MFC=0.548

Table I.9: SVM classifier in the validation set without feature selection

Methods	C	acc	std	fm	std	TPR	std	prec	std
BCorrU	200	**0.632**	0.033	**0.571**	0.038	0.363	0.066	0.475	0.062
BCorrD	200	0.623	0.034	0.562	0.040	0.356	0.068	0.458	0.062
BCohF[0]	35	0.615	0.033	0.520	0.039	0.243	0.055	0.425	0.082
BCohF[1]	90	0.596	0.034	0.522	0.040	0.291	0.066	0.402	0.066
BCohF[2]	50	0.597	0.032	0.519	0.037	0.278	0.061	0.400	0.065
BCohW	1000	0.610	0.032	0.548	0.037	0.343	0.065	0.435	0.058
BH2U	200	0.622	0.034	0.554	0.042	0.333	0.070	0.453	0.067
BH2D	200	0.582	0.036	0.497	0.040	0.245	0.057	0.365	0.073
BMITU	10000	0.612	0.037	0.555	0.042	0.363	0.068	0.441	0.065
BMITD1	10000	0.616	0.037	0.559	0.041	0.366	0.066	0.449	0.063
BMITD2	10000	0.620	0.034	0.561	0.040	0.364	0.066	0.454	0.062
BTEU	500	0.629	0.031	0.525	0.041	0.232	0.060	0.452	0.087
BTED	500	0.624	0.028	0.499	0.040	0.180	0.057	0.423	0.096
				RP_fm=0.488 std=0.043, WRP_acc=0.543 std=0.042, MFC=0.647					

Table I.10: SVM classifier in the test set without feature selection

Methods	C	acc	fm	TPR	prec
BCorrU	200	0.486	0.465	0.318	0.412
BCorrD	200	0.514	0.496	0.364	0.453
BCohF[0]	25	0.507	0.370	0.046	0.250
BCohF[1]	50	0.541	0.436	0.121	0.471
BCohF[2]	1000	0.521	0.496	0.333	0.458
BCohW	150	0.555	0.528	0.348	0.511
BH2U	200	0.534	0.511	0.348	0.479
BH2D	200	0.534	0.478	0.227	0.469
BMITU	10000	0.548	0.525	0.364	0.500
BMITD1	10000	0.568	**0.545**	0.379	0.532
BMITD2	10000	0.521	0.502	0.364	0.462
BTEU	500	0.548	0.498	0.258	0.500
BTED	500	0.521	0.451	0.182	0.429
BCorrU_BCohW	50	0.493	0.477	0.348	0.426
BCorrU_BH2U	30	0.514	0.476	0.273	0.439
BCorrU_BMITD2	50	0.479	0.446	0.258	0.386
BCorrU_BTEU	30	0.507	0.467	0.258	0.425
BCohW_BH2U	100	0.521	0.490	0.303	0.455
BCohW_BMITD2	220	0.555	0.503	0.258	0.515
BCohW_BTEU	250	**0.582**	0.543	0.318	0.568
BH2U_BMITD2	180	0.548	0.507	0.288	0.500
BH2U_BTEU	150	0.500	0.453	0.227	0.405
BMITD2_BTEU	900	0.493	0.437	0.197	0.382
BCorrU_BCohW_BH2U	60	0.534	0.508	0.333	0.478
BCorrU_BCohW_BMITD2	50	0.500	0.473	0.303	0.426
BCorrU_BCohW_BTEU	35	0.534	0.488	0.258	0.472
BCorrU_BH2U_BMITD2	40	0.507	0.462	0.242	0.421
BCorrU_BH2U_BTEU	35	0.514	0.468	0.242	0.432
BCorrU_BMITD2_BTEU	45	0.507	0.462	0.242	0.421
BCohW_BH2U_BMITD2	190	0.507	0.475	0.288	0.432
BCohW_BH2U_BTEU	200	0.562	0.537	0.364	0.522
BCohW_BMITD2_BTEU	250	0.575	0.533	0.303	0.556
BH2U_BMITD2_BTEU	380	0.555	0.517	0.303	0.513
BCorrU_BCohW_BH2U_BMITD2	80	0.521	0.490	0.303	0.455
BCorrU_BCohW_BH2U_BTEU	40	0.521	0.482	0.273	0.450
BCorrU_BCohW_BMITD2_BTEU	70	0.514	0.488	0.318	0.447
BCorrU_BH2U_BMITD2_BTEU	100	0.500	0.476	0.318	0.429
BCohW_BH2U_BMITD2_BTEU	260	0.568	0.542	0.364	0.533
BCorrU_BCohW_BH2U_BMITD2_BTEU	100	0.534	0.511	0.348	0.479
RP_fm=0.498 std=0.043, WRP_acc=0.504 std=0.042, MFC=0.548					

Table I.11: SVM classifier in the validation set with feature selection

Methods	C	acc	std	fm	std	TPR	std	prec	std
BCorrU	42.7	0.637	0.034	0.566	0.038	0.334	0.063	0.484	0.070
BCorrD	1.75	0.642	0.013	0.429	0.025	0.044	0.033	0.552	0.268
BCohF[0]	1.7	0.644	0.011	0.425	0.026	0.039	0.033	0.616	0.293
BCohF[1]	3.45	0.646	0.006	0.419	0.010	0.027	0.009	0.793	0.271
BCohF[2]	4.8	0.640	0.026	0.515	0.049	0.200	0.083	0.508	0.149
BCohW	42.7	**0.647**	0.022	0.479	0.036	0.114	0.046	0.516	0.150
BH2U	8.06	0.609	0.033	0.509	0.040	0.227	0.062	0.406	0.082
BH2D	15.6	0.567	0.041	0.476	0.044	0.224	0.089	0.330	0.080
BMITU	9.14	0.632	0.034	0.567	0.037	0.348	0.062	0.475	0.064
BMITD1	11.9	0.636	0.035	0.569	0.039	0.348	0.065	0.482	0.071
BMITD2	7.2	0.643	0.034	**0.583**	0.040	0.375	0.070	0.496	0.066
BTEU	12.545	0.595	0.039	0.523	0.040	0.295	0.068	0.406	0.072
BTED	31.9	0.598	0.038	0.529	0.040	0.312	0.079	0.412	0.067
RP_fm=0.488 std=0.043, WRP_acc=0.543 std=0.042, MFC=0.647									

Table I.12: SVM classifier in the test set with feature selection

Methods	C	acc	fm	TPR	prec
BCorrU	42.7	0.548	—	0.000	—
BCorrD	1.75	0.562	0.399	0.046	0.750
BCohF[0]	1.7	0.541	0.364	0.015	0.333
BCohF[1]	3.45	0.548	0.392	0.046	0.500
BCohF[2]	4.8	0.555	0.417	0.076	0.556
BCohW	42.7	0.575	0.458	0.121	0.667
BH2U	8.06	0.589	0.539	0.288	0.594
BH2D	15.6	0.562	0.518	0.288	0.528
BMITU	9.14	0.562	0.513	0.273	0.529
BMITD1	11.9	0.589	0.548	0.318	0.583
BMITD2	7.2	0.568	0.551	0.409	0.529
BTEU	12.545	0.596	0.545	0.288	0.613
BTED	31.9	0.418	0.412	0.348	0.354
BCorrU_BCohF[2]	1.7	0.548	—	0.000	—
BCorrU_BH2U	12.85	0.548	—	0.000	—
BCorrU_BMITD2	1.75	0.548	—	0.000	—
BCorrU_BTED	1.7	0.548	—	0.000	—
BCohF[2]_BH2U	1.42	0.534	0.361	0.015	0.250
BCohF[2]_BMITD2	1.7	0.514	0.402	0.091	0.353
BCohF[2]_BTED	1.7	0.500	0.394	0.091	0.316
BH2U_BMITD2	3.3	0.623	0.566	0.288	0.704
BH2U_BTED	19.6	0.658	0.623	0.394	0.722
BMITD2_BTED	9.86	0.603	0.555	0.303	0.625
BCorrU_BCohF[2]_BH2U	1.7	0.548	—	0.000	—
BCorrU_BCohF[2]_BMITD2	32.7	0.562	0.387	0.030	1.000
BCorrU_BCohF[2]_BTED	13.7	0.548	—	0.000	—
BCorrU_BH2U_BMITD2	22.65	0.548	—	0.000	—
BCorrU_BH2U_BTED	32.7	0.548	—	0.000	—
BCorrU_BMITD2_BTED	8.45	0.548	—	0.000	—
BCohF[2]_BH2U_BMITD2	1.6	0.541	0.377	0.030	0.400
BCohF[2]_BH2U_BTED	43.7	0.514	—	0.000	0.000
BCohF[2]_BMITD2_BTED	1.7	0.514	0.402	0.091	0.353
BH2U_BMITD2_BTED	12.9	**0.678**	**0.648**	0.424	0.757
BCorrU_BCohF[2]_BH2U_BMITD2	3.1	0.548	—	0.000	—
BCorrU_BCohF[2]_BH2U_BTED	5.69	0.548	—	0.000	—
BCorrU_BCohF[2]_BMITD2_BTED	1.7	0.555	0.370	0.015	1.000
BCorrU_BH2U_BMITD2_BTED	9.7	0.548	—	0.000	—
BCohF[2]_BH2U_BMITD2_BTED	2.46	0.534	0.373	0.030	0.333
BCorrU_BCohF[2]_BH2U_BMITD2_BTED	1.42	0.548	—	0.000	—

RP_fm=0.498 std=0.043, WRP_acc=0.504 std=0.042, MFC=0.548

Table I.13: RF classifier in the filtered validation set

Methods	est	acc	std	fm	std	TPR	std	prec	std
BCorrU	3	0.557	0.038	0.500	0.040	0.314	0.068	0.357	0.059
BCorrD	7	0.591	0.038	0.508	0.044	0.257	0.065	0.383	0.076
BCohF[0]	3	0.557	0.042	0.497	0.044	0.303	0.070	0.353	0.065
BCohF[1]	3	0.556	0.042	0.496	0.046	0.302	0.073	0.351	0.068
BCohF[2]	3	0.559	0.041	0.499	0.043	0.307	0.072	0.356	0.064
BCohW	3	0.563	0.040	0.507	0.043	0.322	0.070	0.366	0.063
BH2U	3	0.578	0.040	0.524	0.044	0.345	0.073	0.390	0.063
BH2D	3	0.569	0.038	0.513	0.041	0.329	0.070	0.375	0.060
BMITU	5	0.584	0.038	0.516	0.041	0.299	0.066	0.387	0.065
BMITD1	3	0.571	0.038	0.515	0.041	0.331	0.072	0.378	0.060
BMITD2	3	0.574	0.040	0.516	0.043	0.328	0.069	0.382	0.064
BTEU	3	0.590	0.039	0.538	0.043	0.364	0.072	0.411	0.062
BTED	7	**0.607**	0.038	**0.538**	0.045	0.316	0.074	0.424	0.072
RP_fm=0.488 std=0.043, WRP_acc=0.543 std=0.042, MFC=0.647									

Table I.14: RF classifier in the filtered test set

Methods	est	acc	std	fm	std	TPR	std	prec	std
BCorrU	3	0.509	0.041	0.482	0.044	0.317	0.064	0.439	0.066
BCorrD	3	0.519	0.034	0.487	0.038	0.303	0.063	0.451	0.058
BCohF[0]	3	0.522	0.036	0.489	0.039	0.300	0.057	0.457	0.062
BCohF[1]	5	**0.548**	0.035	**0.509**	0.039	0.297	0.058	0.501	0.069
BCohF[2]	3	0.519	0.035	0.487	0.038	0.305	0.061	0.452	0.059
BCohW	3	0.517	0.035	0.482	0.039	0.290	0.061	0.445	0.061
BH2U	3	0.502	0.036	0.475	0.037	0.309	0.056	0.430	0.053
BH2D	3	0.492	0.045	0.467	0.045	0.309	0.060	0.418	0.067
BMITU	3	0.505	0.041	0.481	0.040	0.322	0.055	0.438	0.060
BMITD1	3	0.505	0.039	0.479	0.039	0.318	0.056	0.436	0.057
BMITD2	5	0.510	0.039	0.476	0.042	0.287	0.065	0.437	0.067
BTEU	5	0.520	0.041	0.500	0.044	0.361	0.068	0.460	0.059
BTED	3	0.516	0.042	0.498	0.044	0.370	0.071	0.456	0.060
BCorrD_BCohW	3	0.522	0.036	0.487	0.040	0.294	0.058	0.455	0.063
BCorrD_BH2U	3	0.516	0.042	0.490	0.043	0.325	0.058	0.452	0.066
BCorrD_BMITD2	3	0.510	0.032	0.482	0.036	0.315	0.063	0.440	0.054
BCorrD_BTEU	3	0.511	0.042	0.490	0.043	0.351	0.062	0.448	0.060
BCohW_BH2U	3	0.503	0.036	0.475	0.038	0.304	0.056	0.430	0.055
BCohW_BMITD2	3	0.505	0.036	0.480	0.039	0.323	0.060	0.435	0.055
BCohW_BTEU	3	0.514	0.035	0.495	0.037	0.358	0.062	0.452	0.050
BH2U_BMITD2	3	0.512	0.043	0.486	0.044	0.320	0.059	0.446	0.066
BH2U_BTEU	3	0.514	0.044	0.494	0.045	0.356	0.061	0.453	0.062
BMITD2_BTEU	5	0.527	0.037	0.504	0.037	0.351	0.056	0.471	0.052
BCorrD_BCohW_BH2U	3	0.507	0.038	0.478	0.040	0.305	0.060	0.435	0.060
BCorrD_BCohW_BMITD2	3	0.512	0.042	0.482	0.045	0.310	0.071	0.442	0.066
BCorrD_BCohW_BTEU	3	0.514	0.042	0.492	0.043	0.344	0.062	0.451	0.060
BCorrD_BH2U_BMITD2	3	0.502	0.039	0.474	0.041	0.306	0.060	0.429	0.062
BCorrD_BH2U_BTEU	3	0.514	0.041	0.492	0.045	0.347	0.072	0.449	0.062
BCorrD_BMITD2_BTEU	3	0.519	0.046	0.500	0.047	0.363	0.065	0.460	0.063
BCohW_BH2U_BMITD2	3	0.501	0.040	0.473	0.041	0.305	0.061	0.429	0.063
BCohW_BH2U_BTEU	5	0.527	0.038	0.501	0.041	0.335	0.060	0.466	0.059
BCohW_BMITD2_BTEU	5	0.528	0.037	0.505	0.040	0.349	0.062	0.470	0.058
BH2U_BMITD2_BTEU	5	0.518	0.038	0.495	0.040	0.343	0.065	0.456	0.056
BCorrD_BCohW_BH2U_BMITD2	3	0.509	0.041	0.481	0.043	0.313	0.062	0.440	0.065
BCorrD_BCohW_BH2U_BTEU	3	0.516	0.040	0.493	0.042	0.341	0.061	0.453	0.059
BCorrD_BCohW_BMITD2_BTEU	5	0.521	0.044	0.493	0.046	0.322	0.065	0.458	0.069
BCorrD_BH2U_BMITD2_BTEU	3	0.516	0.044	0.497	0.046	0.361	0.069	0.456	0.062
BCohW_BH2U_BMITD2_BTEU	5	0.521	0.040	0.494	0.041	0.328	0.060	0.459	0.064
BCorrD_BCohW_BH2U_BMITD2_BTEU	3	0.526	0.047	0.506	0.048	0.368	0.066	0.470	0.067
RP_fm=0.498 std=0.043, WRP_acc=0.504 std=0.042, MFC=0.548									

Table I.15: SVM classifier in the filtered validation set without feature selection

Methods	C	acc	std	fm	std	TPR	std	prec	std
BCorrU	200	0.614	0.033	0.533	0.041	0.283	0.066	0.431	0.071
BCorrD	200	0.609	0.035	0.539	0.039	0.314	0.064	0.429	0.066
BCohF[0]	25	0.595	0.028	0.479	0.034	0.177	0.049	0.355	0.077
BCohF[1]	50	0.588	0.033	0.506	0.039	0.260	0.064	0.378	0.067
BCohF[2]	1000	0.586	0.036	0.505	0.042	0.260	0.067	0.376	0.073
BCohW	130	0.594	0.039	0.533	0.042	0.334	0.068	0.409	0.063
BH2U	200	0.604	0.034	0.543	0.040	0.342	0.069	0.424	0.060
BH2D	200	0.581	0.037	0.508	0.041	0.282	0.061	0.377	0.067
BMITU	10000	0.613	0.033	0.566	0.036	0.407	0.068	0.449	0.052
BMITD1	10000	0.614	0.037	0.561	0.041	0.380	0.067	0.447	0.062
BMITD2	10000	**0.631**	0.038	**0.575**	0.042	0.386	0.070	0.474	0.065
BTEU	500	0.599	0.033	0.508	0.039	0.241	0.060	0.393	0.073
BTED	500	0.603	0.031	0.494	0.039	0.201	0.058	0.383	0.083
RP_fm=0.488 std=0.043, WRP_acc=0.543 std=0.042, MFC=0.647									

Table I.16: SVM classifier in the filtered test set without feature selection

Methods	C	acc	fm	TPR	prec
BCorrU	200	0.575	0.544	0.348	0.548
BCorrD	200	0.589	**0.565**	0.394	0.565
BCohF[0]	25	0.582	0.519	0.242	0.593
BCohF[1]	50	0.562	0.537	0.364	0.522
BCohF[2]	1000	0.527	0.491	0.288	0.463
BCohW	150	0.562	0.530	0.333	0.524
BH2U	200	0.527	0.487	0.273	0.462
BH2D	200	0.521	0.473	0.242	0.444
BMITU	10000	0.514	0.488	0.318	0.447
BMITD1	10000	0.500	0.470	0.288	0.422
BMITD2	10000	0.500	0.473	0.303	0.426
BTEU	500	0.555	0.513	0.288	0.514
BTED	500	0.555	0.498	0.242	0.516
BCorrD_BCohW	100	0.596	0.558	0.333	0.595
BCorrD_BH2U	130	0.555	0.513	0.288	0.514
BCorrD_BMITD2	140	0.521	0.486	0.288	0.452
BCorrD_BTEU	80	0.596	0.545	0.288	0.613
BCohW_BH2U	100	0.534	0.472	0.212	0.467
BCohW_BMITD2	180	0.555	0.503	0.258	0.515
BCohW_BTEU	100	0.568	0.497	0.212	0.560
BH2U_BMITD2	180	0.514	0.463	0.227	0.429
BH2U_BTEU	40	0.568	0.503	0.227	0.556
BMITD2_BTEU	60	0.589	0.548	0.318	0.583
BCorrD_BCohW_BH2U	78	0.568	0.503	0.227	0.556
BCorrD_BCohW_BMITD2	117	0.548	0.488	0.227	0.500
BCorrD_BCohW_BTEU	48	0.548	0.476	0.197	0.500
BCorrD_BH2U_BMITD2	105	0.562	0.503	0.242	0.533
BCorrD_BH2U_BTEU	48	0.582	0.524	0.258	0.586
BCorrD_BMITD2_BTEU	56	**0.603**	0.563	0.333	0.611
BCohW_BH2U_BMITD2	110	0.534	0.460	0.182	0.462
BCohW_BH2U_BTEU	58.5	0.555	0.468	0.167	0.524
BCohW_BMITD2_BTEU	70	0.568	0.514	0.258	0.548
BH2U_BMITD2_BTEU	80.5	0.568	0.523	0.288	0.543
BCorrD_BCohW_BH2U_BMITD2	139	0.555	0.503	0.258	0.515
BCorrD_BCohW_BH2U_BTEU	60	0.555	0.475	0.182	0.522
BCorrD_BCohW_BMITD2_BTEU	70	0.568	0.514	0.258	0.548
BCorrD_BH2U_BMITD2_BTEU	80.5	0.596	0.549	0.303	0.606
BCohW[1]_BH2U_BMITD2_BTEU	97	0.534	0.467	0.197	0.464
BCorrD_BCohW_BH2U_BMITD2_BTEU	80.5	0.555	0.487	0.212	0.519

RP_fm=0.498 std=0.043, WRP_acc=0.504 std=0.042, MFC=0.548

Table I.17: SVM classifier in the filtered validation set with feature selection

Methods	C	acc	std	fm	std	TPR	std	prec	std
BCorrU	1.7	0.601	0.037	0.563	0.037	0.436	0.070	0.438	0.052
BCorrD	3.9	0.639	0.027	0.515	0.045	0.197	0.076	0.482	0.105
BCohF[0]	1.63	0.636	0.017	0.433	0.029	0.055	0.040	0.444	0.205
BCohF[1]	1.2	0.647	0.037	**0.594**	0.041	0.409	0.070	0.504	0.064
BCohF[2]	1.68	0.618	0.037	0.499	0.040	0.194	0.075	0.431	0.111
BCohW	1.8	0.609	0.030	0.465	0.038	0.130	0.051	0.357	0.113
BH2U	3.3	**0.650**	0.032	0.547	0.047	0.251	0.074	0.510	0.096
BH2D	56	0.597	0.036	0.509	0.045	0.248	0.067	0.389	0.083
BMITU	6.76	0.614	0.037	0.545	0.041	0.322	0.066	0.440	0.070
BMITD1	5.095	0.612	0.033	0.542	0.038	0.318	0.069	0.434	0.064
BMITD2	2.7	0.625	0.035	0.557	0.040	0.333	0.067	0.461	0.069
BTEU	47	0.602	0.038	0.543	0.042	0.349	0.070	0.424	0.064
BTED	46	0.600	0.040	0.530	0.045	0.309	0.072	0.413	0.073
				RP_fm=0.488 std=0.043, WRP_acc=0.543 std=0.042, MFC=0.647					

Table I.18: SVM classifier in the filtered test set with feature selection

Methods	C	acc	fm	TPR	prec
BCorrU	1.7	0.548	—	0.000	—
BCorrD	3.9	0.596	0.523	0.227	0.652
BCohF[0]	1.63	0.548	0.392	0.046	0.500
BCohF[1]	1.2	0.589	0.562	0.379	0.568
BCohF[2]	1.68	0.521	0.451	0.182	0.429
BCohW	1.8	0.575	0.489	0.182	0.600
BH2U	3.3	0.582	0.513	0.227	0.600
BH2D	56	0.603	0.555	0.303	0.625
BMITU	6.76	0.562	0.498	0.227	0.536
BMITD1	5.095	0.562	0.509	0.258	0.531
BMITD2	2.7	0.555	0.517	0.303	0.513
BTEU	47	0.603	0.567	0.348	0.605
BTED	46	0.603	0.559	0.318	0.618
BCorrU_BCohF[1]	6.8	0.548	—	0.000	—
BCorrU_BH2U	13.8	0.541	0.377	0.030	0.400
BCorrU_BMITD2	7.7	0.527	—	0.000	0.000
BCorrU_BTEU	8.8	0.548	—	0.000	—
BCohF[1]_BH2U	2.8	0.610	0.576	0.364	0.615
BCohF[1]_BMITD2	4.5	**0.616**	**0.589**	0.394	0.619
BCohF[1]_BTEU	10.8	0.603	0.574	0.379	0.595
BH2U_BMITD2	11.275	0.541	0.465	0.182	0.480
BH2U_BTEU	18.1	0.555	0.481	0.197	0.520
BMITD2_BTEU	18.3	0.555	0.517	0.303	0.513
BCorrU_BCohF[1]_BH2U	11.3	0.548	—	0.000	—
BCorrU_BCohF[1]_BMITD2	19.8	0.548	—	0.000	—
BCorrU_BCohF[1]_BTEU	4.915	0.548	—	0.000	—
BCorrU_BH2U_BMITD2	89.8	0.575	0.489	0.182	0.600
BCorrU_BH2U_BTEU	13.425	0.548	0.380	0.030	0.500
BCorrU_BMITD2_BTEU	5.46	0.534	—	0.000	0.000
BCohF[1]_BH2U_BMITD2	39.15	0.603	0.571	0.364	0.600
BCohF[1]_BH2U_BTEU	1.8	0.596	0.558	0.333	0.595
BCohF[1]_BMITD2_BTEU	12.4	0.596	0.565	0.364	0.585
BH2U_BMITD2_BTEU	4.625	0.562	0.486	0.197	0.542
BCorrU_BCohF[1]_BH2U_BMITD2	2.02	0.555	0.370	0.015	1.000
BCorrU_BCohF[1]_BH2U_BTEU	7.85	0.548	—	0.000	—
BCorrU_BCohF[1]_BMITD2_BTEU	11.8	0.548	—	0.000	—
BCorrU_BH2U_BMITD2_BTEU	16.7	0.575	0.474	0.152	0.625
BCohF[1]_BH2U_BMITD2_BTEU	9.4	0.575	0.541	0.333	0.550
BCorrU_BCohF[1]_BH2U_BMITD2_BTEU	23.6	0.562	0.399	0.046	0.750
RP_fm=0.498 std=0.043, WRP_acc=0.504 std=0.042, MFC=0.548					