UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

The effect of using multiple connectivity metrics in brain Functional Connectivity studies

Hugo Emanuel Augusto Teixeira

Mestrado Integrado em Engenharia Biomédica e Biofísica
Perfil em Sinais e Imagens Médicas

Dissertação orientada por:
Professor Doutor Alexandre Andrade

2021

Acknowledgements

Firstly, I would like to thank the Faculdade de Ciências da Universidade de Lisboa, for integrating me so well over these years. I would also like to thank the teachers I had the opportunity to meet, for always transmitting the necessary knowledge in the best way possible and for being always close to the students.

I would also like to express my gratitude to my advisor, Professor Doctor Alexandre Andrade from the Faculdade de Ciências da Universidade de Lisboa and Instituto de Biofísica e Engenharia Biomédica, for all the shared knowledge that was truly important for this dissertation, confidence in the work developed throughout these months, and for the opportunity to work in such an interesting area, leading to my personal and academic growth.

To all my friends and colleagues, I'm extremely grateful for always being there for me over these years, whether to help me with academic or other matters, and for all the great moments we've been through.

A special mention to Marta, for being the best person I know, for always supporting and motivating me even when things didn't seem to go well, for all the valuable opinions, and for helping me becoming the best version of myself.

Last but not least, there are no words that can express the unconditional support, motivation, and love that my family, especially my mother and grandparents, have given me throughout my life and my academic journey. They are the best, I wouldn't be the person I am today and I wouldn't have achieved what I've achieved without them, a thank you is not enough.

Abstract

Resting-state functional magnetic resonance imaging (rs-fMRI) has the potential to assist as a diagnostic or prognostic tool for a diverse set of neurological and neuropsychiatric disorders, which are often difficult to differentiate. fMRI focuses on the study of the brain functional Connectome, which is characterized by the functional connections and neuronal activity among different brain regions, also interpreted as communications between pairs of regions. This Functional Connectivity (FC) is quantified through the statistical dependences between brain regions' blood-oxygen-level-dependent (BOLD) signals time-series, being traditionally evaluated by correlation coefficient metrics and represented as FC matrices. However, several studies underlined limitations regarding the use of correlation metrics to fully capture information from these signals, leading investigators towards different statistical metrics that would fill those shortcomings. Recently, investigators have turned their attention to Deep Learning (DL) models, outperforming traditional Machine Learning (ML) techniques due to their ability to automatically extract relevant information from high-dimensional data, like FC data, using these models with rs-fMRI data to improve diagnostic predictions, as well as to understand pathological patterns in functional Connectome, that can lead to the discovery of new biomarkers. In spite of very encouraging performances, the black-box nature of DL algorithms makes difficult to know which input information led the model to a certain prediction, restricting its use in clinical settings.

The objective of this dissertation is to exploit the power of DL models, understanding how FC matrices created from different statistical metrics can provide information about the brain FC, beyond the conventionally used correlation family. Two publicly available datasets where studied, the ABIDE-I dataset, composed by healthy and autism spectrum disease (ASD) individuals, and the ADHD-200 dataset, with typically developed controls and individuals with attention-deficit/hyperactive disorder (ADHD). The computation of the FC matrices of both datasets, using different statistical metrics, was performed in MATLAB using MULAN's toolbox functions, encompassing the correlation coefficient, non-linear correlation coefficient, mutual information, coherence and transfer entropy. The classification of FC data was performed using two DL models, the improved ConnectomeCNN model and the innovative ConnectomeCNN-Autoencoder model. Moreover, another goal is to study the effect of a multi-metric approach in classification performances, combining multiple FC matrices computed from the different statistical metrics used, as well as to study the use of Explainable Artificial Intelligence (XAI) techniques, namely Layer-wise Relevance Propagation method (LRP), to surpass the black-box problem of DL models used, in order to reveal the most important brain regions in ADHD.

The results show that the use of other statistical metrics to compute FC matrices can be a useful complement to the traditional correlation metric methods for the classification between healthy subjects and subjects diagnosed with ADHD and ASD. Namely, non-linear metrics like h^2 and mutual information, achieved similar and, in some cases, even slightly better performances than correlation methods. The use of FC multi-metric, despite not showing improvements in classification performance compared to the best individual method, presented promising results, namely the ability of this approach to select the best features from all the FC matrices combined, achieving a similar performance in relation to the best individual metric in each of the evaluation measures of the model, leading to a more complete classification. The LRP analysis applied to ADHD-200 dataset proved to be promising, identifying brain regions related to the pathophysiology of ADHD, which are in broad accordance with FC and structural study's findings.

Keywords: Functional Connectivity, Connectome, Brain Disorders, Deep Neural Networks, Neural Networks Explainability

Resumo

A ressonância magnética funcional em estado de repouso (rs-fMRI) tem o potencial de ser uma ferramenta auxiliar de diagnóstico ou prognóstico para um conjunto diversificado de distúrbios neurológicos e neuropsiquiátricos, que muitas vezes são difíceis de diferenciar. A análise de dados de rs-fMRI recorre muitas vezes ao conceito de conectoma funcional do cérebro, que se caracteriza pelas conexões funcionais entre as diferentes regiões do cérebro, sendo estas conexões interpretadas como comunicações entre diferentes pares de regiões cerebrais. Esta conectividade funcional é quantificada através de dependências estatísticas entre os sinais fMRI das regiões cerebrais, sendo estas tradicionalmente calculadas através da métrica coeficiente de correlação, e representadas através de matrizes de conectividade funcional. No entanto, vários estudos demonstraram limitações em relação ao uso de métricas de correlação, em que estas não conseguem capturar por completo todas as informações presentes nesses sinais, levando os investigadores à procura de diferentes métricas estatísticas que pudessem preencher essas lacunas na obtenção de informações mais completas desses sinais.

O estudo destes distúrbios neurológicos e neuropsiquiátricos começou por se basear em técnicas como mapeamento paramétrico estatístico, no contexto de estudos de fMRI baseados em tarefas. Porém, essas técnicas apresentam certas limitações, nomeadamente a suposição de que cada região cerebral atua de forma independente, o que não corresponde ao conhecimento atual sobre o funcionamento do cérebro. O surgimento da rs-fMRI permitiu obter uma perspetiva mais global e deu origem a uma vasta literatura sobre o efeito de patologias nos padrões de conetividade em repouso, incluindo tentativas de diagnóstico automatizado com base em biomarcadores extraídos dos conectomas. Nos últimos anos, os investigadores voltaram a sua atenção para técnicas de diferentes ramos de Inteligência Artificial, mais propriamente para os algoritmos de *Deep Learning* (DL), uma vez que são capazes de superar os algoritmos tradicionais de *Machine Learning* (ML), que foram aplicados a estes estudos numa fase inicial, devido à sua capacidade de extrair automaticamente informações relevantes de dados de alta dimensão, como é o caso dos dados de conectividade funcional. Esses modelos utilizam os dados obtidos da rs-fMRI para melhorar as previsões de diagnóstico em relação às técnicas usadas atualmente em termos de precisão e rapidez, bem como para compreender melhor os padrões patológicos nas conexões funcionais destes distúrbios, podendo levar à descoberta de novos biomarcadores. Apesar do notável desempenho destes modelos, a arquitetura natural em caixa-preta dos algoritmos de DL, torna difícil saber quais as informações dos dados de entrada que levaram o modelo a executar uma determinada previsão, podendo este utilizar informações erradas dos dados para alcançar uma dada inferência, restringindo o seu uso em ambientes clínicos.

O objetivo desta dissertação, desenvolvida no Instituto de Biofísica e Engenharia Biomédica, é explorar o poder dos modelos DL, de forma a avaliar até que ponto matrizes de conectividade funcional criadas a partir de diferentes métricas estatísticas podem fornecer mais informações sobre a conectividade funcional do cérebro, para além das métricas de correlação convencionalmente usadas neste tipo de estudos. Foram estudados dois conjuntos de dados bastante utilizados em estudos de Neurociência e que estão disponíveis publicamente: o conjunto de dados ABIDE-I, composto por indivíduos saudáveis e indivíduos com doenças do espectro do autismo (ASD), e o conjunto de dados ADHD-200, com controlos tipicamente desenvolvidos e indivíduos com transtorno do défice de atenção e hiperatividade (ADHD).

Numa primeira fase foi realizada a computação das matrizes de conetividade funcional de ambos os conjuntos de dados, usando as diferentes métricas estatísticas. Para isso, foi desenvolvido código de MATLAB, onde se utilizam as séries temporais dos sinais BOLD obtidas dos dois conjuntos de dados

para criar essas mesmas matrizes de conectividade funcional, incorporando funções de diferentes métricas estatísticas da caixa de ferramentas MULAN, compreendendo o coeficiente de correlação, o coeficiente de correlação não linear, a informação mútua, a coerência e a entropia de transferência. De seguida, a classificação dos dados de conectividade funcional, de forma a avaliar o efeito do uso de diferentes métricas estatísticas para a criação de matrizes de conectividade funcional na discriminação de sujeitos saudáveis e patológicos, foi realizada usando dois modelos de DL. O modelo ConnectomeCNN melhorado e o modelo inovador ConnectomeCNN-Autoencoder foram desenvolvidos com recurso à biblioteca de Redes Neuronais Keras, juntamente com o seu *backend* Tensorflow, ambos em Python. Estes modelos, desenvolvidos previamente no Instituto de Biofísica e Engenharia Biomédica, tiveram de ser otimizados de forma a obter a melhor *performance*, onde vários parâmetros dos modelos e do respetivo treino dos mesmos foram testados para os dados a estudar. Pretendeu-se também estudar o efeito de uma abordagem multi-métrica nas tarefas de classificação dos sujeitos de ambos os conjuntos de dados, sendo que, para estudar essa abordagem as diferentes matrizes calculadas a partir das diferentes métricas estatísticas utilizadas, foram combinadas, sendo usados os mesmos modelos que foram aplicados às matrizes de conectividade funcional de cada métrica estatística individualmente. É importante realçar que na abordagem multi-métrica também foi realizada a otimização dos parâmetros dos modelos utilizados e do respetivo treino, de modo a conseguir a melhor *performance* dos mesmos para estes dados. Para além destes dois objetivos, estudou-se o uso de técnicas de Inteligência Artificial Explicável (XAI), mais especificamente o método *Layer-wise Relevance Propagation* (LRP), com vista a superar o problema da caixa-preta dos modelos de DL, com a finalidade de explicar como é que os modelos estão a utilizar os dados de entrada para realizar uma dada previsão. O método LRP foi aplicado aos dois modelos utilizados anteriormente, usando como dados de entrada o conjunto de dados ADHD-200, permitindo assim revelar quais as regiões cerebrais mais importantes no que toca a um diagnóstico relacionado com o ADHD.

Os resultados obtidos mostram que o uso de outras métricas estatísticas para criar as matrizes de Conectividade Funcional podem ser um complemento bastante útil às métricas estatísticas tradicionalmente utilizadas para a classificação entre indivíduos saudáveis e indivíduos como ASD e ADHD. Nomeadamente métricas estatísticas não lineares como o h^2 e a informação mútua, obtiveram desempenhos semelhantes e, em alguns casos, desempenhos ligeiramente melhores em relação aos desempenhos obtidos por métodos de correlação, convencionalmente usados nestes estudos de conectividade funcional. A utilização da multi-métrica de conectividade funcional, apesar de não apresentar melhorias no desempenho geral da classificação em relação ao melhor método das matrizes de conectividade funcional individuais do conjunto de métricas estatísticas abordadas, apresenta resultados que justificam a exploração mais aprofundada deste tipo de abordagem, de forma a compreender melhor a complementaridade das métricas e a melhor maneira de as utilizar. O uso do método LRP aplicado ao conjunto de dados do ADHD-200 mostrou a sua aplicabilidade a este tipo de estudos e a modelos de DL, identificando as regiões cerebrais mais relacionadas à fisiopatologia do diagnóstico do ADHD que são compatíveis com o que é reportado por diversos estudos de conectividade funcional e estudos de alterações estruturais associados a esta doença. O facto destas técnicas de XAI demonstrarem como é que os modelos de DL estão a usar os dados de entrada para efetuar as previsões, pode significar uma mais rápida e aceite adoção destes algoritmos em ambientes clínicos. Estas técnicas podem auxiliar o diagnóstico e prognóstico destes distúrbios neurológicos e neuropsiquiátricos, que são na maioria das vezes difíceis de diferenciar, permitindo aos médicos adquirirem um conhecimento em relação à previsão realizada e poder explicar a mesma aos seus pacientes.

Palavras-chave: Conetividade Funcional, Conectoma, Distúrbios Cerebrais, Redes Neuronais Profundas, Explicabilidade de Redes Neuronais

Contents

List of Figures

List of Tables

Acronyms

AAL	Automated Anatomical Labelling
ABIDE-I	Autism Brain Imaging Data Exchange I
ADHD	Attention-deficit/hyperactive disorder
AI	Artificial Intelligence
AD	Alzheimer's disease
ASD	Autism spectrum disorder
AUC	Area under the ROC curve
BOLD	Blood oxygen level dependent
BRAPH	Brain Analysis using Graph Theory
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DAN	Dorsal attention network
DMN	Default mode network
DL	Deep Learning
DNN	Deep Neural Network
DW-MRI	Diffusion-weighted Magnetic Resonance Imaging
DTI	Diffusion Tensor Imaging
EC	Effective Connectivity
EEG	Electroencephalography
FC	Functional Connectivity
fMRI	Functional Magnetic Resonance Imaging
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
freqs	Frequencies
fs	Sampling Frequency

GPU	Graphics Processing Unit
h^2	Non-linear correlation coefficient
Hz	Hertz
LRP	Layer-wise Relevance Propagation
MAE	Mean Absolute Error
MSE	Mean Squared Error
MEG	Magnetoencephalography
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MULAN	Multiple Connectivity Analysis
NPV	Negative Positive Value
rs-fMRI	Resting-state functional magnetic resonance imaging
ROI	Region of Interest
ReLU	Rectified linear unit
ROC	Receiver Operating Characteristics
SC	Structural Connectivity
SZ	Schizophrenia
SELU	Scaled Exponential Linear Units
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
TR	Repetition Time
Tanh	Tangent
VAN	Ventral attention network
XAI	Explainable Artificial Intelligence
2D	Two-dimensional
3D	Three-dimensional

1 – Introduction

1.1 – Context and Motivation

The brain is the most complex organ in the human body, being composed by an estimate of 8.6×10^{11} neurons that connect with each other via approximately 10^{14} synapses, allowing chemical and electrical signals to be transmitted, either by efferent or afferent pathways [1,2]. The human brain is considered a very efficient network, with a large number of functionally and structurally interconnected regions that are specialized to perform certain functions and are constantly sharing information with each other [3].

The rs-fMRI became one of the most used techniques to study the brain since the study performed by Biswal et al. (1995), reporting highly correlated spontaneous activity from right and left motor cortices when a subject was at rest, showing that brain activity is present even in the absence of specific tasks [4]. The advances in non-invasive Neuroimaging techniques and brain network analysis paved the way to a new the field in Neuroscience, the brain Connectome, linking the structural and functional information of the brain network based on the idea of a brain circuit map. This consists of brain regions with their structural connections and respective functional interactions, allowing to know the behavior of the system as a whole and the interactions between different regions of the brain, both at a structural and functional level [3,5]. Researchers have started to explore the human brain network from the perspective of connectivity patterns, with much of its attention being focused on the study of FC, which helps characterize not only healthy individuals but several neurological and neuropsychiatric disorders, like schizophrenia (SZ), ASD, Alzheimer disease (AD) and ADHD, being defined by measure the relationship between BOLD signals from distinct regions of the brain using statistical metrics, traditionally correlation.

For many years, the study of neurological and neuropsychiatric disorders through FC, has relied on mass-univariate analytical techniques like statistical parametric mapping, which compared healthy patients with disease diagnosed patients to report neuroanatomical and neurofunctional differences, providing significant improvements towards the understanding of these disorders. However, these techniques have limitations, such as the statistical inferences assuming that each brain region acts independently, which is not true, and only allow to detect differences between groups. Partly to fulfill those limitations, but also with the goal of improving diagnostic power, neuroimaging researchers began to focus on Machine Learning (ML) algorithms, a branch of Artificial Intelligence (AI) responsible for extracting patterns from data and learning how to make predictions in new data [6]. With increasingly improved ML techniques, these began to be used in the study of FC of the brain, overcoming the drawbacks existing with mass-univariate analytical methods, enhancing the diagnosis capacity and knowledge regarding FC patterns in neurological and neuropsychiatric disorders, achieving varying degrees of success [7,8].

Despite their vast and popular use, conventional ML algorithms lacked good performance on raw data, requiring the use of expertise to extract of the most important features, which ends up being extremely arduous due to the complex high-dimensional datasets from rs-fMRI and FC data, with the attention turned to the application of another branch of AI called DL. DL algorithms are representation-learning methods, meaning that they can automatically extract and learn good representations from the raw data, without the need of manual feature selection as with conventional ML algorithms. Beside that quality, DL models have the ability to reach higher levels of complexity and abstraction, making them perfectly suited to the classification of complex FC data [7]. Researchers began to take advantage of the potential of DL models and apply them to FC data, where several studies showed their promising results

in FC based classification of neurological and neuropsychiatric disorders [6,8]. The use of DL models can be extremely valuable in the analysis of diverse neurological and neuropsychiatric disorders, as these are difficult to differentiate since the diagnosis is based on clinical interview to determine signs and symptoms present, with symptoms being shared between diseases, as well as subtle neuroanatomical and neurofunctional abnormalities.

Although having a superb performance, DL models lack an explanation about which characteristics of the input features were used to achieve a given outcome. To fulfill this limitation, several XAI techniques were developed to provide transparency to these models, explaining and evaluating which input data features the models are using to achieve a given prediction [9]. The field of XAI techniques is still taking its first steps in its application to medical data, but it has enormous potential for the incorporation of these AI models in clinical environments, allowing to bring machine logic closer to clinicians, also facilitating the presentation of the results of this to the patients.

1.2 - Objectives

After a brief introduction about the context of this work, there are several goals to be accomplished during this dissertation project:

1. Use of multiple statistical metrics to compute FC matrices, beyond the traditional use of correlation, and assessment of the performance of each individual metric by using automatic classifiers, mainly DL models, in order to classify between healthy and diseased subjects.

2. Understand if coupling together those connectivity metrics used to compute the FC matrices, creating a multi-metric, can indeed improve the performance of DL models used previously to distinguish subjects from a diseased or healthy state.

3. Evaluate which input features, in other words brain regions, from the FC matrices computed using the set of statistical metrics, are relevant and positively contribute to the classification and discrimination of ADHD subjects, by using the XAI technique LRP. Following this, compare whether the relevant brain regions obtained by using LRP are in line with what has been reported by previous studies.

1.4 – Scientific Contribution to this Dissertation

For this dissertation, in order to achieve the objectives proposed to accomplish, it was created a self-developed code in MATLAB to compute the FC matrices through the BOLD time-series data retrieved from the online databases ABIDE-I and ADHD-200, incorporating the statistical metrics functions from MULAN toolbox. To study the classification performance of the FC matrices from the brain disorders addressed, two DL models were used and optimized, the ConnectomeCNN and ConnectomeCNN-Autoencoder models, which were previously developed by the researcher Antonio Cano Montes at the Instituto de Biofísica e Engenharia Biomédica. To conclude the final goal of this dissertation, an XAI technique called LRP, from the Python's iNNvestigate toolbox, has been coupled up to the DL models used to help understand how these models are using the FC data to perform a given prediction and identify which regions of the brain are more associated with an ADHD-related diagnosis.

1.5 – Dissertation Outline

This dissertation project is divided into six chapters. The first chapter is responsible for providing a general overview of this dissertation project thematic, what led to its development and the main objectives to be achieved. The chapter 2 focuses on the general concepts involving this work, where it

includes notions about data acquisition with rs-fMRI, how the FC matrices are created using different metrics and concepts about Deep Neural Networks (DNNs), mainly Convolutional Neural Networks (CNNs) and autoencoders, how they work, their development and how to evaluate their performances. Still on this chapter, it is introduced the XAI technique called LRP, which is used to unveil the black-box problem present in DL models. Chapter 3 approaches the state-of-the-art regarding the Connectome field and what are the main limitations nowadays. Chapter 4 describes the methodology used in this project, as well as the materials needed for its development. In Chapter 5, the results obtained are showed, followed by their respective discussion. The final chapter seals this project by highlighting the main findings and conclusions, some limitations of the study and guidelines for future research.

2 – Theoretical Background

2.1 – Resting-state Functional MRI

Magnetic Resonance Imaging (MRI) is a non-invasive technique introduced at the clinical stage in 1980s. MRI is an extraordinary versatile imaging technique able to provide high-resolution images, both two-dimensional (2D) and three-dimensional (3D) data from different human body structures, since muscles, cartilage, organs, white matter tracts and arteries. This technique, because of the absence of ionizing electromagnetic radiation, has become increasingly used in many clinical applications such as diagnosis, staging and treatment monitoring of a wide range of pathologies, including neurological and neuropsychiatric disorders [10].

The basic principle on which MRI works is the magnetizing properties of the atomic nuclei, where an external magnetic field is applied through the patient to align the protons, that are randomly oriented in the water nuclei of the examined location, with that field. This alignment is the consequence of the magnetization effect, which will be further disturbed by the application of an external radiofrequency energy pulse, causing these nuclei to return to their normal alignment through relaxation processes, emitting radiofrequency energy. Then, moments after the application of radiofrequency energy, the emitted signals are measured using the Fourier transformation, which translates the frequency information of these signals into the respective intensity levels, of each location in the image plane, being displayed as grayscale in a matrix composed of these pixels. Different radiofrequency energy pulses can be generated and received, in order to create different types of MRI images [10].

The fMRI is a non-invasive modality of MRI that, instead of analyzing the anatomical structure of the brain, it examines brain activity by observing the neurological processes, regional or time-varying changes that influence the brain metabolism consumption [1]. The fMRI is performed under the same principles as the conventional MRI scans, where it uses Nuclear Magnetic Resonance coupled with gradients in magnetic field to create images of the patient's brain structures neuronal activity, with that neuronal activity being based on blood oxygen level dependent (BOLD) signal, that is the basis of fMRI formation, being totally dependent on the oxygen levels in the blood, which are influenced by the brain metabolic activity [5,11].

Blood contains oxyhemoglobin and deoxyhemoglobin, which are diamagnetic and paramagnetic molecules, respectively. The deoxyhemoglobin present in a blood vessel leads to a susceptibility between the vessel and the neighboring tissue, causing the dephasing of MR proton signal and a darkening of the image in the regions containing those vessels, with the oxyhemoglobin, as diamagnetic molecule, will not cause the same dephasing. It would be expected that with the enhancement of neuronal activity, the concentration of deoxyhemoglobin would consequently increase due to the consumption of oxygen and decrease the signal. Instead, associated with the increase of oxygen levels from the blood is the increase in cerebral blood flow, which transport with it more oxyhemoglobin, observed in figure 2.1, reducing the concentration of deoxyhemoglobin and increasing the BOLD signal [12]. The BOLD signal changes present some smooth delays or lags from the beginning of the neuronal activity, characterizing the changes in blood flow that is detected by the fMRI, with this delay being called hemodynamic response. The deoxyhemoglobin involved in BOLD signal can be affected by many factors, such as changes in cerebral blood flow and volume, cerebral metabolic rates of oxygen and different magnetic fields strength [5,11].

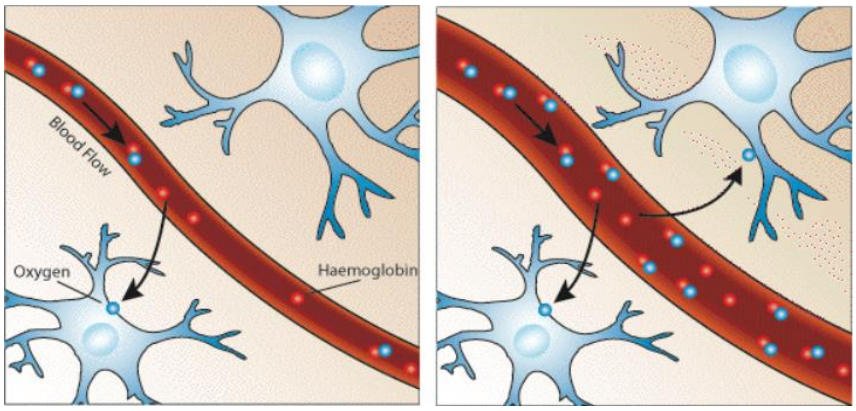


Figure 2.1: Neuronal activity increases oxygen levels and consequently cerebral blood flow, leading to higher levels of oxyhemoglobin [4].

During many years, in task-based fMRI studies, the standard procedure, it was thought that the presence of some oscillations in BOLD signals recorded was noise resulted from physiological processes like cardiac pulsation, respiratory and subject movement, leading to the rejection of that "noise" from the main signal to posterior analysis. But, as demonstrated by Biswal et al [3], part of this problematic "noise" present in the signals was in fact the so-called brain spontaneous fluctuations, which refers to activity that is not originated from specific stimulus towards the patients, representing neuronal activity intrinsically generated by the human brain, as illustrated in figure 2.2 [4]. These spontaneous fluctuations are consistent low frequency fluctuations, in the order of 0.01-0.08 Hertz (Hz), which will be considered in rs-fMRI and are confined to distinct cortical network systems in the brain [3,11].

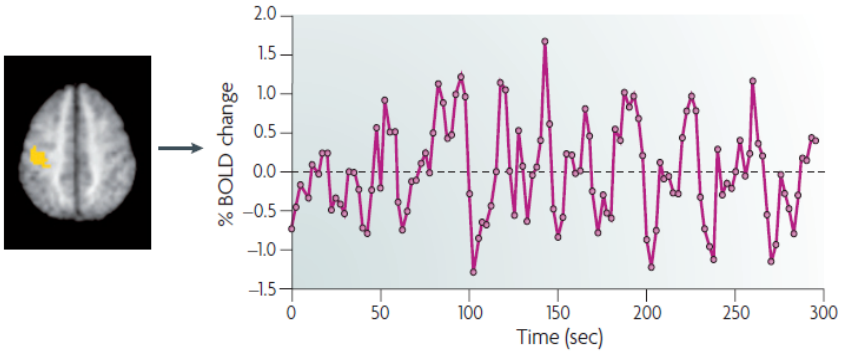


Figure 2.2: Representation of resting-state blood oxygen level dependent signal activity from a brain region [4].

These spontaneous fluctuations in rs-fMRI demonstrate that the human brain is always operational and working, representing at rest about 20% of the total energy consumption of the body, most of that energy is used to support the spontaneous neuronal signals that are taking place. The energy consumption of the brain when performing tasks is 5% less than the 20% of resting-state body's total energy consumption, thus showing the importance of the resting-state in brain functions, providing a window to be a disease-related signal change [4,13]. When two brain regions show a highly correlated BOLD signal during the rs-fMRI, they are said to be functionally correlated, even if those regions are not structurally connected by any direct pathway [14].

The wide use of fMRI in comparison to other techniques is mainly due to its excellent spatial resolution, which allows the measurement of BOLD signal changes, while techniques like Electroencephalography (EEG) and Magnetoencephalography (MEG) have a poor spatial resolution and excellent temporal resolution. Another advantage of fMRI in comparison to other techniques is the

capability to detect deeper brain activity changes and it offers a better signal-to-noise ratio. The advantages of using resting-state in fMRI are the easy and short period of acquisition, allowing an increased sample size, and unlike task-based imaging, resting-state allows the observation of many brain networks at once [11,15]. Since is not necessary the execution of a task, rs-fMRI circumvents the confuse interpretation of tasks, allowing greater comfort and less effort from the patients. Even though rs-fMRI has all these advantages compared to task-based fMRI, these signals are influenced by many physiological processes, such as the cardiac pulsation, respiratory and subject movements, which are not related to the neuronal activity, requiring preprocessing procedures to be later analyzed.

2.2 – Brain Connectivity

The human brain is a hierarchical complex that comprises different yet connected levels, from genes, proteins, synapses, neurons and their circuits, brain areas and their pathways, and the brain as a whole. The concept that the human brain was a complex, large-scale network, called Connectome, emerged in 2005 by Sporns et al, although the idea that the brain was a structural network of connections between neurons with functional implications, had been proposed by Santiago Ramón y Cajal decades earlier [16]. Today, the definition of brain Connectome is based on a mapping of the brain circuit, consisting in brain regions, their structural connections, and respective functional interactions, allowing to understand the dynamic interactions between different brain regions, both at a structural and functional level [17,18]. There are different types of connectivity that provide information to study the Connectome, these being the Structural Connectivity (SC), the FC and the Effective Connectivity (EC).

The brain structural Connectome consists of grey matter, representing the neuronal elements where information is processed, and white matter tracts, which will be the structure where communication pathways rely on [19]. The SC is based on an anatomic map of physical connections comprising white matter fiber tracts, linking different brain cortical and subcortical regions with their fiber bundle [20]. Neuronal axons involving these white matter fiber tracts allow them to transmit neural signals to other brain areas, which is fundamental for the communication between them [21]. To assess SC, techniques are available including Diffusion-weighted MRI (DW-MRI), left image in figure 2.5, a variant from conventional MRI, which is very sensitive to the water diffusion within brain tissues, measuring the magnitude of diffusion for each tissue voxel, generating a contrast map based on comparisons between the differences in water diffusion values in brain tissues [22,23]. Diffusion Tensor Imaging (DTI) can also be used to study SC, middle image in figure 2.3, providing images of anisotropy of water diffusion in the brain, offering information about its structure and direction of diffusion, because myelin creates a barrier to water diffusion and white matter tracts show a substantial diffusion anisotropy [24]. Together with DTI can be used the Tractography, right image in figure 2.5, which uses the orientation information to estimate the structural pathways between brain regions and reconstruct the direction of axons tracts, with the SC being accomplished by calculate the number of streamlines in a certain pair of brain regions [21,22].

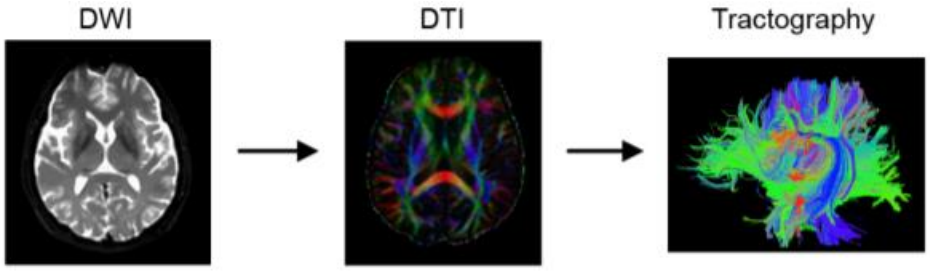


Figure 2.3: Procedure used to study Structural Connectivity depicted by white matter fiber tracts [25].

Another type of brain connectivity is the FC, defined as the synchronization and patterns of interactions between different brain regions that do not necessarily need to be structurally connected and may result from direct anatomical connections or remote paths [8,26]. FC is characterized by the measure of temporal correlations or statistical dependences among time-series of BOLD signals between different brain regions, as illustrated in figure 2.4, being indicative of neural activity over time, at each voxel [1,20]. Functional communication between brain regions is very important in many complex processes, assisting in the continuous integration of information from different brain regions, making this connectivity highly important in the comprehension of human brain organization and disorders patterns [3]. These types of connections can oscillate on small time intervals such as seconds or milliseconds, being time dependent and focusing on spontaneous brain activity of ongoing information processing between regions [16,17].

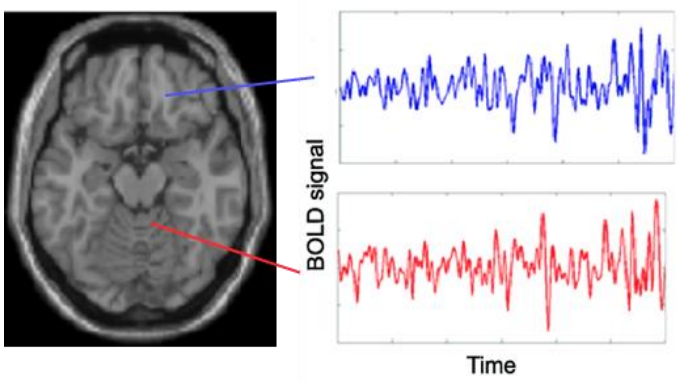


Figure 2.4: Procedure to study Functional Connectivity between two brain regions. Adapted from [27].

Several studies suggest that disruptions in FC are the root of several brain disorders, making this type of connectivity more vulnerable to pathologies [2]. FC has been an important tool to examine how brain organization and functional connections might be changed in neurological and psychiatric disorders, contributing to an earlier and very useful diagnosis of these disorders [3,16]. Closely related and linked to functional brain interactions, emerges the EC, which estimates the influence that a neural element exerts over another, evaluating the directionality and the causality of neural interactions [1,11]. Both FC and EC are derived from the relationships between brain regions BOLD signal time-series, which can be acquired with fMRI, EEG or MEG, but resting-state fMRI is achieving a higher use in this research due to the properties mentioned previously [20].

2.2.1 – Functional Connectivity Metrics

As mentioned previously, FC employs statistical methods to evaluate the neuronal dependencies between different brain regions BOLD signal time-series, called in this dissertation as FC metrics. There is a significant set of metrics that can be used in FC analysis, being those subdivided in different categories, according to their functions and mathematical formulations. The first subdivision that can be made is related to the domain where the metrics are applied, as they can be in time domain or in frequency domain. Another subdivision is based on whether the metric considers linear or non-linear dependencies between the signals, being called linear FC metrics or non-linear FC metrics. Another subdivision is focused on the capacity of the metric to quantify the direction of interaction between regions signals, in other words, the objective is to understand which region causes the effects in the other, as it can be directed or non-directed metric [28].

Correlation

The most widely used metric to evaluate the dependency between neuronal signals is the correlation, as it is simple to calculate and being widely used in FC studies facilitates the exchange of knowledge between researchers [29]. If one region of brain is functionally connected to another, even though they are distant, should be present correlation regarding their BOLD signal time-series. This metric, given in equation 2.1, by considering two brain regions time-series signals *x* and *y*, allows to calculate the linear correlation among pairs of brain regions, being the covariance between the signals time-series cov_{xy}, dividing by the product of both signals' standard deviations, σ_x and σ_y [30,31]. The calculation of correlation coefficient has an important variable corresponding to the time delay or lag of the hemodynamic response present in BOLD signal, the τ, which in the case of Pearson's correlation coefficient, a correlation family metric, is considered as zero [32].

$$R_{xy}(\tau) = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}} = \frac{cov_{xy}(\tau)}{\sigma_x . \sigma_y} \quad (2.1)$$

Correlation values range from -1 to +1, where -1 indicates that as the value of one signal increases, the value of the other signal decreases, being negatively correlated, and +1 indicates that as the value of one signal increases, the value of the other signal does the same, being positively correlated. It is important to mention that this metric do not provide information about the direction or causality between the signals compared [31].

Coherence

Another linear metric used in FC studies is the coherence, which is equivalent to correlation but is computed from the frequency domain of the signals. This undirected metric allows to overcome the problems of artifact noises such as cardiac or respiratory activity, which could result in high illusory correlations [32]. Fourier-based coherence or simply coherence, defined by the equation 2.2, measures the common energy between pairs of signals at a specific frequency, quantifying the amount of variance in one of the signals that can be explained by the other [28,33].

$$coh_{xy}(f) = \frac{\left|F_{xy}(f)\right|^2}{F_{xx}(f) . F_{yy}(f)} \quad (2.2)$$

The mathematical basis of this metric consists of F_{xy}, which is the cross-spectrum between signals *x* and *y* at a frequency *f*, defined by the Fourier transform of the cross variance in equation 2.3, and of F_{xx} and F_{yy} being the power-spectrum of the signals *x* and *y* time-series at a frequency *f*, respectively calculated through equations 2.4 and 2.5 [28,33].

$$F_{xy}(f) = \sum_u cov_{xy}(u) . e^{-jfu} \quad (2.3)$$

$$F_{xx}(f) = \sum_u cov_{xx}(u) . e^{-jfu} \quad (2.4)$$

$$F_{yy}(f) = \sum_u cov_{yy}(u) . e^{-jfu} \quad (2.5)$$

Due to the non-stationarity changes in neuronal signals time-series, their spectral characteristics vary over time, coherence should be considered over the time domain as well. Derived from that concept, beside the Fourier-based coherence mentioned, emerged the Wavelet-based coherence (*Wcoh*), which

corresponds to the measure of correlation between a pair of signals in time-frequency domain [34]. Signals can be decomposed through the Morlet wavelet family, despite the existence of a variety of wavelet functions, this one is the most used because of its simplicity and is well suited for spectral estimations, having a good stability between time and frequency. Morlet wavelet (Ψ), mathematically described in equation 2.6, is defined for both frequency *f* and time τ, being the product of sinusoidal wave at a certain frequency with a Gaussian function centered at a certain time and with the standard deviation (σ) proportional to the inverse of frequency [34,35].

$$\Psi_{\tau,f}(u) = \sqrt{f} \times e^{i2\pi f(u-\tau)}.e^{-\frac{(u-\tau)^2}{\sigma^2}} (2.6)$$

From the convolution of a signal *x* with the Morlet wavelet is obtained the wavelet transform of that respective signal, as a function of time τ and frequency *f*, where * denotes the complex conjugate:

$$W_x(\tau,f) = \int_{-\infty}^{+\infty} x(u).\Psi_{\tau,f} * (u)du \ (2.7)$$

Finally, the calculation of Wavelet-based coherence, defined by equation 2.9, will be equivalent to the Fourier-based coherence introduced in equation 2.3, by using the wavelet cross-spectrum between a signals *x* and *y* time-series in both time and frequency domain, provided by the equation 2.8, and using the product of power-spectrum from each signal around time and frequency [35].

$$F_{xy}(\tau,f) = \int_{\tau+\frac{\delta}{2}}^{\tau+\frac{\delta}{2}} W_x(\tau,f).W_y(\tau,f)\,d\tau (2.8)$$

$$Wcoh(\tau,f) = \frac{\left|F_{xy}(\tau,f)\right|^2}{F_{xx}(\tau,f).F_{yy}(\tau,f)} \ (2.9)$$

Both Fourier-based and Wavelet-based coherence have values ranging from 0 to +1, indicating that the signals have no linear relationship, and that one signal can predict the other in a linear way, respectively. The use of coherence is particularly interesting, as this metric allows to the study the dependencies between neuronal signals in the range of low frequencies, where the spontaneous fluctuations of the brain occur and are closely related to the FC, namely between 0.01-0.08 Hz, as previously mentioned [32].

Non-linear correlation coefficient

In addition to the linear FC metrics mentioned so far, non-linear metrics are also used to study the dependency between neuronal signals from brain regions. Among the non-linear statistical analysis methods, Lopes da Silva et al. [36] developed a metric originally for EEG signals analysis and has been applied to the field of brain FC analysis in recent years, describing the dependency of a signal *x* on another signal *y*, independently of the relation between these two signals. This is based on the idea that if the value of a signal *x* is considered as a function of a value from signal *y*, the given value from signal *x* can be predicted by means of a non-linear regression curve [37].

This metric developed is represented as non-linear correlation coefficient (h^2), basing itself in a scatter plot between signals *y* and *x*, with the signal *x* to be split into bins and the mean value of signal *y*, as well as signal *x* value of the midpoint, being calculated for each respective bin. Through the connection of the points previously calculated, an approximation of the regression curve is achieved and

the h^2 is calculated as follows in equation 2.10, where the $f(x_i)$ is the linear piecewise approximation of the non-linear regression curve [36,37].

$$h^2_{y|x} = \frac{\sum_{k=1}^{N} y(k)^2 - \sum_{k=1}^{N} (y(k) - f(x_i))^2}{\sum_{k=1}^{N} y(k)^2} \quad (2.10)$$

The values obtained for h^2 range from 0, where the two signals are independent from each other, to +1, when one signal is fully determined by the other. Differing from the traditional correlation, which is always symmetric, the h^2 ratio can be asymmetric, where the relationship explained from signal *x* to signal *y* may be different to the relationship from signal *y* to *x*, with the amount of asymmetry being related to the nature of the respective relationship [38].

Mutual Information

Another non-linear approach used to study the relationship between neuronal signals time-series is mutual information, which is based on concepts from information theory [45]. Information theory was developed with the aim of measure the entropy of a random variable, which is the amount of information or uncertainty required to specify the outcome of that variable, that is known or can be estimated [39].

The information content present in a random variable *x* can be explored through Shannon entropy, defined by equation 2.11, which consists in splitting the signal into *M* bins and represent the probability density p_i^x that a measurement will find *x* in the i^{th} element of the bin, being represented through a histogram of the respective bin, where the sum is extended to all the values that the variable can assume [30,40].

$$H(x) = -\sum_{i \in M} p_i^x . ln\, p_i^x \quad (2.11)$$

The definition of Shannon entropy can be extended to a pair of multivariate random variables *x* and *y*. As stablished by equation 2.12, the variables are divided into *M* bins and it is used the joint entropy of those variables, which is defined according to the variable's joint probability density p_{ij}^{xy}, instead of each signal probability, involving finding the values of the random variables in two different spaces, M_x and M_y [30,40].

$$H(x,y) = -\sum_{i \in M_x} \sum_{j \in M_y} p_{ij}^{xy} . ln\, p_{ij}^{xy} \quad (2.12)$$

The notion of information theory from Shannon entropy is expanded to characterize mutual information, which is a statistical method that quantifies the overlap of the information content present between time-series from two signals, being the reduction of uncertainty of one signal due to the knowledge of the other. One of the properties present in mutual information is that its calculation between signals *x* and *y*, is exactly the same applied between signals *y* and *x*, being thus symmetrical [40]. Differentiating from correlation, mutual information is an undirected metric that can measure both linear and non-linear relationships between time-series, being susceptible to dependences that are not exhibited in the covariance [41]. In terms of its mathematical formulation, mutual information assumes the form of equation 2.13, where M_x and M_y are the potential values that *x* and *y* can take, with p_{ij}^{xy} being the probability that the signals take the values *i* in M_x and *j* in M_y, and p_i^x and p_j^y assuming the probability functions of each signal [45].

$$MI(x,y) = H(x) + H(y) - H(x,y) = \sum_{i \in M_x} \sum_{j \in M_y} p_{ij}^{xy} . ln \frac{p_{ij}^{xy}}{p_i^x p_j^y} \quad (2.13)$$

If the result of mutual information is 0, the measurement of a value from signal *x* time-series is totally independent of the measurement of a value from signal *y* time-series, having no information shared between these signals. On the other hand, if mutual information result is +1, reaches its maximum value, where the two signals time-series are completely the same [40].

Transfer Entropy

As seen previously, mutual information has the ability to provide evidence about the amount of shared information content between two signals time-series but, it says little about existent causal interactions, due to the shortage of directional and dynamical information [42]. Despite some use of Granger causality to study causal relationships in time-series data, it is limited to linear interactions, making it inadequate to study causal relationships in highly complex non-linear systems like human brain [42,43].

Also based in information theory emerged transfer entropy, which is a direct and non-linear statistical measure that quantifies the reduction of uncertainty in the future values of a signal *y* by knowing the past and present values of signal *x* instead of only knowing the past and present values of signal *y*. Thomas Schreiber introduced the concept of transfer entropy in 2000, allowing to estimate the amount of information flow from a signal to another, with its definition based on principle of observational causality from the mathematician Norbert Wiener, where a signal *x* is said to cause a signal *y* when the next value of signal *y* is better predicted by knowing the past and present of signal *x* than using the past and present of signal *y* alone [44]. When two signals *x* and *y* can be approximated by Markov processes, Schreiber defined a measure of causality from the generalized Markov condition in equation 2.14, where $x_t^m = (x_t, \ldots, x_{t-m+1})$ and $y_t^m = (y_t, \ldots, y_{t-n+1})$ [45].

$$p(y_{t+1} \mid y_t^n, x_t^m) = p(y_{t+1} \mid y_t^n) \quad (2.14)$$

Transfer entropy can be understood by the conditional mutual information, presented in equation 2.15, which allows to describe the information transfer between two signals by considering the history of both signals, x_t^m and y_t^n, with the parameters *m* and *n* being the number of states considered from the past of each respective signal [44]. The state parameters include the most important past observations of the respective signal over time [45]. The conditioning in equation 2.15 enables transfer entropy to fulfill the drawbacks present in mutual information, by allowing to assess the directional and dynamical information between the two signals, due to its asymmetric property, where transfer entropy from signal *x* to signal *y* is not the same as the transfer entropy from signal *y* to signal *x*, and because it is based on transition properties between states, respectively [44,45].

$$TE_{x \to y} = MI(Y_{t+1}; X_t^m \mid Y_t^n) \quad (2.15)$$

For two time-series *x* and *y*, Schreiber uses the Kullback-Leibler divergence between the two distributions at each side of equation 2.14 to mathematically define transfer entropy, as shown in equation 2.16, where x_t^m and y_t^n are, as before, the history of both signals time-series for the respective states *m* and *n*, while y_{t+1} refers to the state of time-series *y* at a time *t* + 1. Further, $p(y_{t+1}, y_t^n, x_t^m)$ is denoted as the joint probability of y_{t+1} and the histories x_t^m and y_t^n of the two time-series, with $p(y_{t+1} \mid y_t^n, x_t^m)$ and $p(y_{t+1} \mid y_t^n)$ representing the conditional probabilities [43,46].

$$TE_{x\to y} = \sum_{y_{t+1}} \sum_{y_t^n} \sum_{x_t^m} p(y_{t+1}, y_t^n, x_t^m) \,.\, ln\left(\frac{p(y_{t+1} \mid y_t^n, x_t^m)}{p(y_{t+1} \mid y_t^n)}\right) \quad (2.16)$$

It is very important to highlight that transfer entropy captures causal dependencies through some value in the past of a signal to explain the future of another signal, further than the past of the latter, causing transfer entropy to actually obtain the knowledge about information transfer between signals instead of quantifying the strength of causal relationships [44].

2.2.2 – Brain Network Analysis

To study the FC among different regions of the brain, as is the scope of this study, the rs-fMRI BOLD signals must go through some preprocessing, since these signals are exposed to artifacts that are not related to the neural activity recorded, influencing the nature of the signals. The artifacts that influence BOLD signals are mainly physiological processes like cardiac pulsation, respiratory and subject movements, being these identified and removed in preprocessing steps [47]. After preprocessing, two more procedures are executed in order to perform an analysis of the brain network study, these being brain regions definition and FC measurement.

Human brain is a complex network with a huge number of neurons, so reconstructing the entire Connectome at that scale and for the variability of existing brains is a difficult task. The process of spatial partitioning of the brain into macroscale regions is called brain parcellation, helping in the reduction of information from thousands of voxels into a group of nodes and to select the regions of interest (ROIs) [48,49]. In brain parcellation representations each parcel is responsible for a node in the network, subdividing different brain regions, as demonstrated in figure 2.5C, with the number of regions in the parcellation to be considered, as they play an important role in estimating further characteristics of the network [50]. To select the nodes of the parcellated brain network, the most widely used option in connectivity studies is the Automated Anatomical Labelling (AAL) atlas, where the AAL atlas using 116 brain regions is the most commonly applied, with the brain being parcellated by using a pre-defined anatomical template that is human-crafted, based essentially on cytoarchitectural characteristics [51]. There are other pre-defined anatomical atlases options like the Harvard-Oxford atlas and the Desikan–Killiany atlas, with these having individual characteristics such as the number and spatial location of brain regions, as well as the image registration technique used. Beside the pre-defined anatomical atlases mentioned, atlases can be generated from random-voxel seeds, in which a voxel in the gray matter or a small triangle on the gray-white matter boundary surface can be used as a node, producing random equal size dividing parcels that divide the brain regions uniformly [48,50].

Once the brain is subdivided into different parcels at a macroscale level, reducing the complexity of the brain, the extraction of BOLD signal time-series from the *N* parcels of the atlas chosen is made, as seen in figure 2.5D, and employed to create the FC matrix, which will then be used to study the functional connections and statistical dependencies between brain regions. The FC matrix establishes the relation between the time-series from all nodes/parcels of a network, where a 2D array with *N* rows and *N* columns is created, with each row and respective column representing a unique node from the network. The $N \times N$ matrix is fulfilled by the values obtained from the computation of the FC metrics, described in subchapter 2.2.1, between every pair of nodes time-series, resulting in a connectivity matrix similar to the one represented in figure 2.5E, with the values range depending on the statistical metric used to estimate the connectivity.

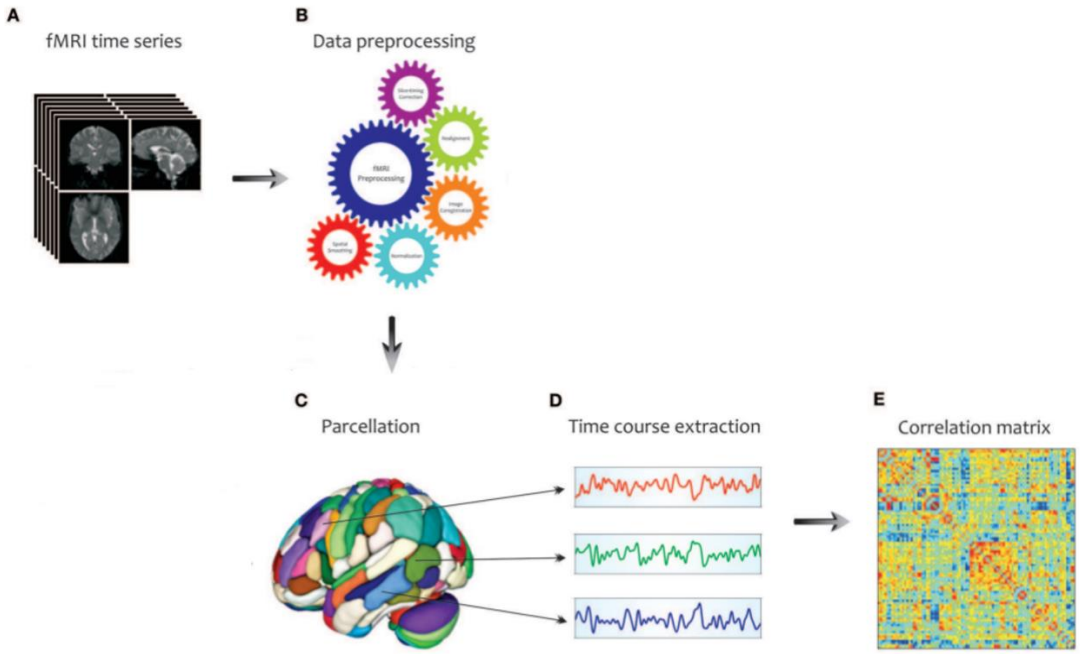


Figure 2.5: Representation of steps involved in brain connectivity analysis using Functional Magnetic Resonance Imaging data. Adapted from [1].

This FC matrix has become an important component of research in many Neuroscience investigation studies, mostly using fMRI, helping researchers to understand the network-level properties of the brain, how tasks can reconfigure the brain and how the dysfunctions in those networks, mainly due to neurological diseases, propagate and affect the relationships among different brain regions, allowing to find patterns and information that permits to differentiate patients, including making a diagnosis and prognosis of these neurological disorders [52]. One of the emerging techniques to use FC matrix data to diagnose neurological disorders is a subfield of ML, called DL.

2.3 – Deep Learning

With the combined evolution of human knowledge and technology, the process of having machines that could have the ability of learn without being manually programmed became possible, emerging the field of AI and ML. ML uses algorithms to teach machines how to interpret and learn information from the data provided, being used in innumerous applications, including Neuroimaging studies [53]. The learning of a ML machine can be supervised, unsupervised, semi-supervised or by reinforcement. Supervised learning is a task of learning where the model uses the input data to achieve an output, by being trained on already labeled/targeted data. This labeled data corresponds to a dataset that includes the inputs and the expected outputs, with the algorithm finding methods to determine how to reach those outcomes from the inputs. In unsupervised learning, unlike supervised learning, there are no labels to aid the learning from the model, with the machine having the responsibility to discover important features from the available data. Semi-supervised learning is a combination of both learning methods, with a portion of the dataset having the corresponding labels and the remaining data is unlabeled, where the algorithm can learn how to predict the unlabeled data from the data already labeled. The reinforcement learning uses as example the learning experience from humans, by trial and error, the machine explores different options and possibilities, in terms of parameters and actions, learning from each result and evaluate which one is optimal [54].

The performance of ML algorithms depends on good quality of the input data, where a bad quality of data can lead to a lower performance, requiring careful handcrafted feature engineering to transform the input data in its raw format into learnable data, so that ML models are be able to identify patterns in the input for further classification [53,55]. With the expansion and availability of data, feature engineering turned out to be too arduous to keep up. To tackle the drawbacks created by ML models, a new branch of ML methods emerged, allowing machines to use the raw input data and automatically distinguish patterns/features for classification, the so-called DL models or DNNs, inspired by Neural Networks, mimicking the functioning processes of human brain [53]. DL is one of the most exciting research topics in many fields, especially in Neuroscience, undergoing major advances in solving various problems that have persisted, by taking advantage of the increasing amount of data available and computational resources.

2.3.1 – Concepts of Neural Networks

The history of the creation of a system that could resemble the human brain function started in 1943 with Warren McCulloch and Walter Pitts, as they were trying to understand how human brain could be so complexly interconnected by their basic cells neurons, creating the McCulloch and Pitts model of a neuron. This was an extremely important contribution that paved the way for Frank Rosenblatt, in 1958, create the first prototype of a Neural Network, called the perceptron [56].

As observed in figure 2.6, the perceptron is a single-layer Neural Network, due to having one layer linking the input and output, is used in supervised learning to distinguish the data between two classes and is composed by four important parts including input values, weights and bias, the weighted sum and the activation function. Weights are values that can be adjusted for the network be trained to accomplish a desired output and contain the knowledge of the Neural Network about the problem, where a positive weight value represents a strong connection of that input regarding the result and a negative weight value represents the opposite. The inputs are multiplied by their respective weight and are combined to create the weighted sum. Bias is a special type of input that is used to adjust the output along with the weighted sum, allowing to shift the activation function to right or left and to have higher quality in training.

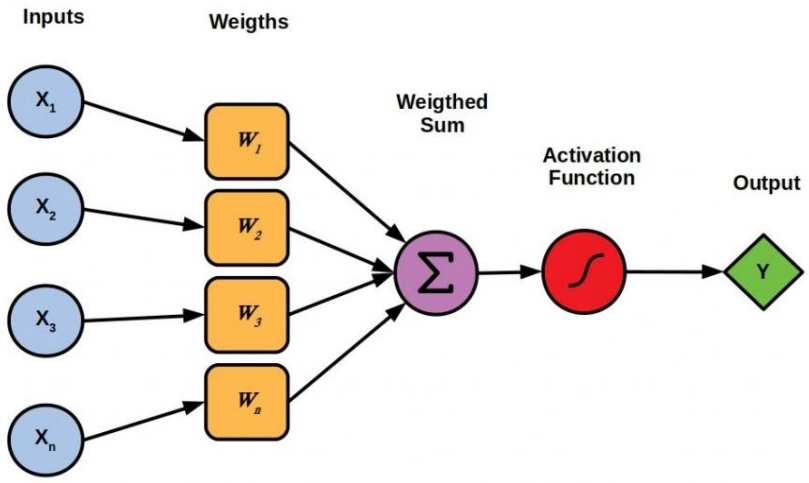


Figure 2.6: Structure of the perceptron [57].

Mathematically, the perceptron is defined by equation 2.17, where y will be the outcome predicted, x_i refers to all the inputs used from a dataset and w_i the weights of those respective inputs, with b being the bias applied to the network and a the activation function.

$$y = f(x) = \left(\sum_{i=1}^{m} x_i w_i + b \right) a \quad (2.17)$$

Activation functions introduce non-linearities to the network, using the values from the weighted sum to perform mathematical operations to convert them into interpretable values for the classification process. One of the most used activation functions is the sigmoid function, which takes an input value and outputs a value between 0 and +1, but it has some disadvantages like saturation and vanishing of the gradients, responsible for the correct update in the direction and quantity of the network weights, and the output values are not zero-centered, causing the gradients to oscillate between positive and negative values. Another activation function used is the hyperbolic tangent (Tanh) function, with output values varying from -1 to +1, and despite not having the problem of zero-centered outputs, the saturation of gradients remains [58]. Due to the problems presented by these two activation functions, emerged the rectified linear unit (ReLU) function. ReLU will output the respective input value if this one is positive, while if the input is negative, it will output zero. With its linearity overcomes the vanishing gradient problem in other activation functions, allowing models to learn faster and perform better [58,59].

Despite being able to learn from the data and execute predictions, the perceptron by having one adaptive layer is limited to only recognize linearly separable patterns, blocking its application in more complex tasks. This limitation fell off with the introduction of backpropagation or multilayer perceptron networks, the basis for DNNs algorithms, which extends the original perceptron by adding multiple hidden layers between the inputs and output. A typical multilayer perceptron contains the input, output and hidden layers between the input and output, and non-linear computational elements called neurons or units, as seen in figure 2.7, with the neurons of one layer are fully connected to neurons in adjacent layers. Hidden layers of the network have the ability to identify and extract features present in the input data, which helps resolving more complex problems than the original perceptron. This capacity arises from the internal mappings of input data patterns that occur in hidden layers during training, which posteriorly uses those mapped features of the input to automatically recognize them in the classification phase [60].

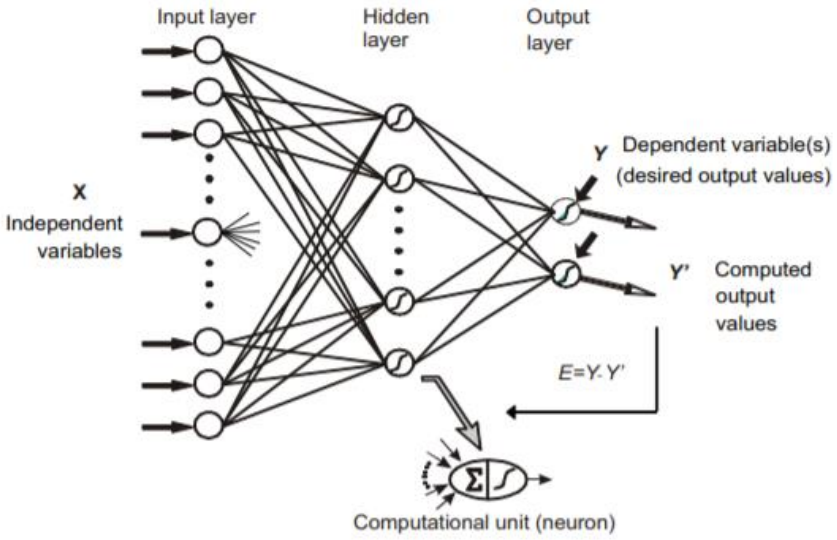


Figure 2.7: Structure of a multilayer perceptron with an input and output layer, plus one hidden layer in between [60].

The multilayer perceptron computation for a two-layer network is given by equation 2.18, where is similar to the equation 2.17 of perceptron, with the only difference being the number of calculations

performed, instead of being limited to input-output layers, extends to input-hidden layers and then to hidden-output layers. Here, y is the predicted output as in the perceptron, W_1 and W_2 represent the weights of the first and second layers, respectively, $b_1$and b_2 are the biases applied to the first and second layers, respectively, with $a_1$and a_2 being the activation functions of each layer.

$$y = f(x) = a_2(a_1\ (x \cdot W_1 + b_1) \cdot W_2 + b_2)\ \ (2.18)$$

2.3.1.1 – Training, Optimization and Shortcomings of Neural Networks

The learning is the main part of DNNs, being an optimization process, where the model is trained to find the best parameters (weights) in the network that minimize the loss or cost function, the error between the classification output computed by the model and the desired target values. There are some loss functions used to determine the error, among them the Mean Squared Error (MSE) and Mean Absolute Error (MAE). The learning of a Neural Network involves two steps: a forward-propagation step followed by a backpropagation step.

Forward-propagation starts by feeding the input layer with a given set of input data, which will go through the network and each input is multiplied by their respective weights, with the weights being randomly initialized as small numbers. The forward-propagation continues with the activation function calculations with the weighted sum, sum of the multiplication of all inputs by their weights, which will propagate forward through the hidden layers, where in each following layer the previous process is repeated, until it reaches the output layer of the network, producing the predicted classification value for the original input [60]. After the estimated classification outcome, the error or loss function for each output, the difference between the desired target and the network output, are calculated. The amount of error is then backpropagated from the output layer towards the input layer, being used to update the weights into new ones, with the objective of produce outputs closer to the desired target values, reducing the loss function.

To know how to minimize the error obtained between the expected output and the prediction of the network in order to find the optimal values for the weights towards a better output, it is necessary an optimization function called gradient descent. Batch gradient descent is one of the most used type of gradient descent in ML and DL, where the weights matrix is randomly initialized and then runs through all the input data used to train the model before update model's network weights by calculate the gradient of loss function. Another widely used gradient descent is a variation of batch gradient descent called stochastic gradient descent, differing in the fact that the weights updates are made after running over *n* number of random samples from the input data used to train the network, allowing a faster convergence when compared with gradient descent, meaning that the network has rapidly and successfully learned how to respond to the patterns of the data [59].

In this training learning process, the speed at which the weights are updated is managed by a model hyperparameter called learning rate. When the learning rate is too low, it takes too much time to find an optimal state, while higher learning rate values will reduce the loss faster but incorrectly. In order to determine the best learning rates, emerged the adaptive learning algorithms, allowing to adapt the learning rate in response to certain parameters [55]. Several adaptive algorithms have been proposed over the years to tackle the limitations of the adaptive gradient algorithm (AdaGrad), the first optimization method developed, with the adaptive moment estimation (Adam) being one of the most popular and with better performance among all. Adam combines the ideas of AdaGrad, root mean square propagation (RMSProp) and momentum, other adaptive learning algorithms, providing adaptive learning rates for each parameter. Adam retain the exponential decaying average of past squared

gradients v_t, as made by AdaGrad and RMSProp, as well as retaining the exponentially decaying average of past gradients m_t, just like momentum [59].

Firstly, the decaying averages of past and past squared gradients are calculated through equations 2.19 and 2.20, where β_1 and β_2 are hyperparameters known as decay rates that control the contribution of past recorded gradients versus the actual gradient, respectively, with g_t being the vector of gradients for the current iteration.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (2.19)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (2.20)$$

The m_t and v_t are estimations of the first and second moments of the gradients, respectively, which are then used in Adam weights update rule described by equation 2.21, with θ_t as the neuron weight for an iteration *t*, η as the learning rate and ϵ assumes a small value to prevent divisions by zero.

$$\theta_{t-1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \quad (2.21)$$

The purpose of training the network is to make sure that the model has successfully acquired the best knowledge to perform well in unseen data from the dataset used and others, but it's not always the case. One of the main shortcomings of Neural Networks is undoubtedly the overfitting problem, which is characterized by the loss of the model's ability to generalize to other data than the data used to train [60]. When a network is overfitted, it cannot learn general patterns present in the training data, but learns instead specific characteristics of that training data, causing the error between the expected output and the outcome predicted to increase, as seen in figure 2.8.

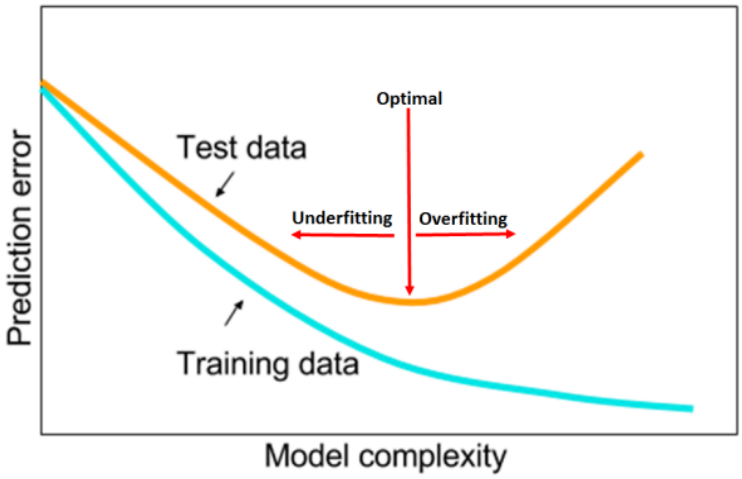


Figure 2.8: Representation of the differences between good fitting, overfitting and underfitting [61].

In order to reduce overfitting, there are several options that can be used, starting with the addition of more data. If there is more data available from the dataset used, it should be added, as the performance of Neural Networks is significantly improved with the increase of examples provided. In case of not being possible to add more data, there is the technique of data augmentation, which takes the selected data to train the network and applies a series of operations, like rotation, translation, or size changes in image data types, for example.

One technique successfully used to tackle overfitting is the dropout regularization. In dropout, some neurons in a layer are randomly selected to have their activation ceased during forward-propagation and backpropagation, having their weights and outputs set to zero. Each time the training is executed, different sets of neurons are dropped, preventing the adaptation of the network to the training data used. Another technique used is the regularization, where the model is optimized by applying penalties to complex models, forcing the network to be simple and reduce the loss function. Dropout performs better than regularization in reducing the overfitting, improving the training speed as well [59].

There are hyperparameters in the models that can contribute to the problem of overfitting, such as learning rate, batch size, number of data samples used to later update the network parameters, and epochs, number of iterations specified to train the model and go through the number of data samples chosen in batch size. The determination of batch size and epochs values influence the learning rate, which consequently influence the ability of the network to converge and find optimal solutions, which may lead to overfitting of the model [60]. A solution for this is the use of a technique called early stopping, which allows to indicate a threshold value to stop the training process when the model performance is no longer improving, by using for example a predefined value for loss function.

2.3.2 – Convolutional Neural Networks

Different DNNs architectures were developed to tackle various problems in several areas, with CNNs getting a lot of attention, having a great success specially in computer vision and the detection, segmentation and recognition of objects and regions in images, including in Connectome [53]. Its structure is similar to traditional Neural Networks because it is inspired by neurons in the human and animal brain, more specifically by the visual cortex.

The architecture of CNNs is structured by three main types of neural layers, each one with different functions in the network, designed to process data arranged in 3D, $m \times m \times r$, with m referring to the height and width of the input, and r to the number of channels, 3 for an RGB (red, green and blue) image and 1 for a grayscale image. The model starts with convolutional layers, the main component of CNNs and where most of the computation occurs. Each convolutional layer is composed by filters or kernels k of size $n \times n \times q$, with n being smaller than the input image and q equal to or smaller than r, which are extremely important as they are convolved with small regions of the original input data, together with the respective weights and bias (W and b), to produce feature maps (h), which contain information about the input data [55]. The convolution operation is represented in figure 2.9.

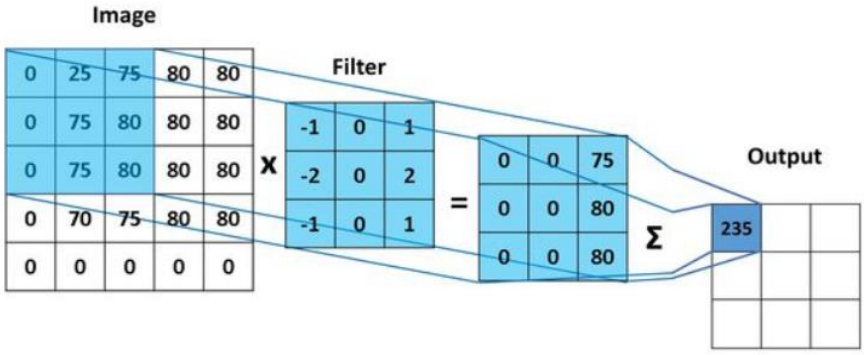


Figure 2.9: Convolution operation in convolutional layers [62].

Mathematically speaking, the convolutional layer computes a dot product between inputs x and respective weights W, followed by the addition of a bias b, with an activation function f then applied to the output of the product, as demonstrated by equation 2.22 [55].

$$h^k = f\left(W^k * x + b^k\right) \text{ (2.22)}$$

In convolutional operation, filters have an important parameter called stride, which is a used to specify the size of the movement made by the filter through the input, vertically and horizontally. The size of strides directly affects the output volume obtained, where a bigger stride results in smaller output volumes, as it performs the convolution operation with more input data points. Stride parameter works together with padding, which is a parameter that adds empty units to the data in order to cover more input data, resulting in more input information, as observable in bottom image sequence of figure 2.10, and, consequently, in more accurate analysis [58]. With this link, stride and padding can be used to adjust the dimensionality of the data when using convolutional layers.

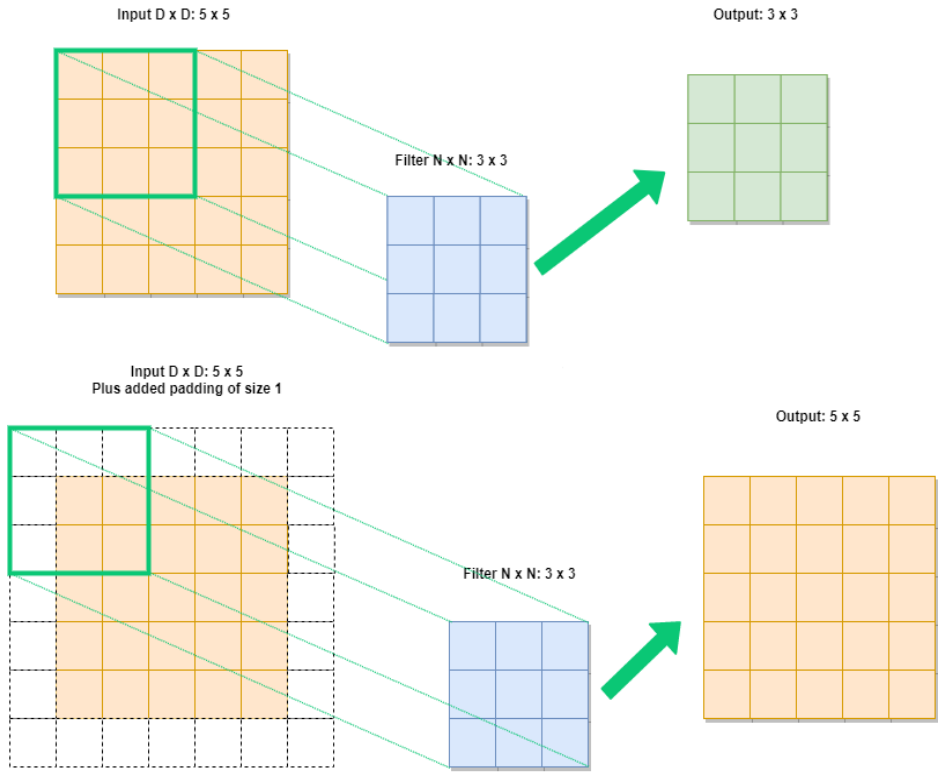


Figure 2.10: Comparison of padding types, with the top image sequence having no padding added, while the image sequence below has a size 1 padding addition. Adapted from [63].

Alternating with convolutional layers are pooling layers, where the feature maps are subsampled to decrease the number of parameters in the model by reducing the width and height of the input, leading to a loss of information, being helpful to accelerate the training and control the problem of overfitting. Pooling operations are performed over a $p \times p$ region for all feature maps, with p as the filter size used to run over the feature map. There are different pooling operations, with the most used being max pooling and average pooling, where in the first one maintains the highest value present in the $p \times p$ region, while in the other the statistical mean of the values present in the $p \times p$ region is applied. Of these two, max pooling can converge faster, select superior invariant features and improve generalization of the model [56].

Following the set of convolutional and pooling layers, as final layers are the fully connected layers, just like the one used in multilayer perceptron, that take the features generated previously and create high-level of abstraction from the data. Neurons in fully connected layers are fully connected to all the activation of the previous layers, converting then the features maps into one-dimensional feature vector,

which will be fed into the last layer to obtain the classification scores, where each score is the probability of a given instance corresponding to a certain class [55,56]. It is important to highlight that the learning backpropagation in CNNs is the same as other DNNs and the multilayer perceptron, allowing all the weights in all the filters to be trained [53]. In figure 2.11 is an example of the construction of a CNN architecture, from its input to the output.

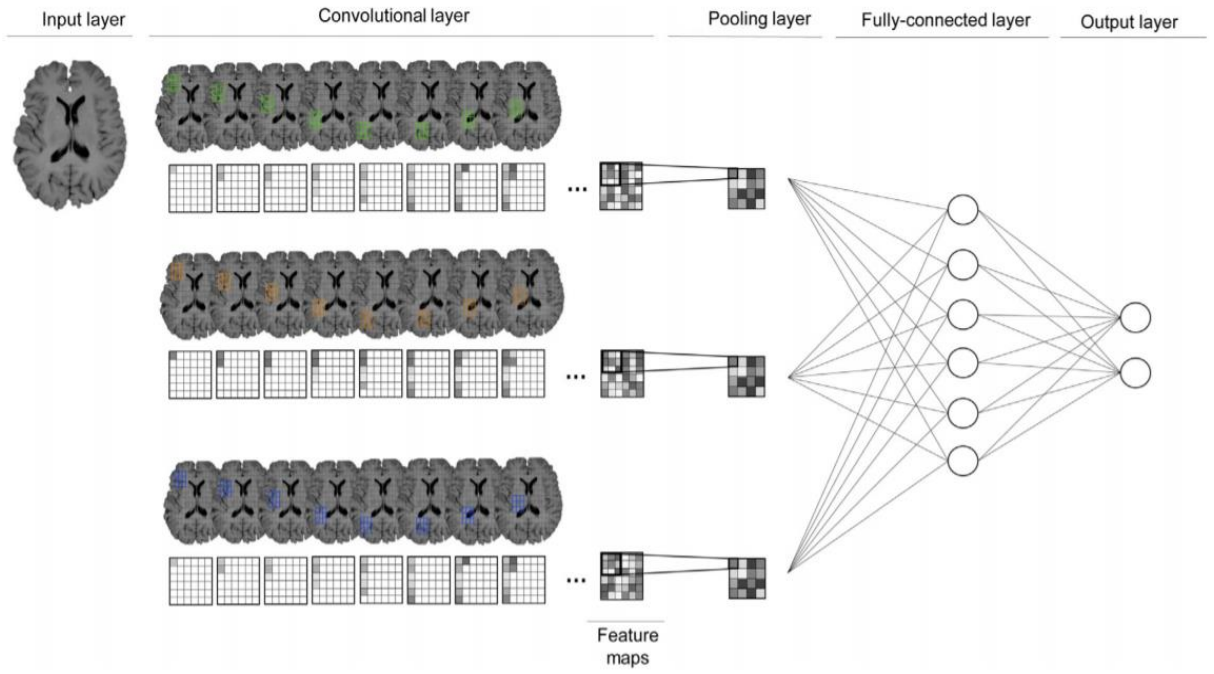


Figure 2.11: Example architecture of a basic Convolutional Neural Network [6].

2.3.3 –Autoencoders

As mentioned previously, for a conventional ML to perform well, a good feature representation from the dataset is required, which needs specialized expertise, leading to a time consuming and very difficult task. The increasingly need for the development of algorithms that were capable of automatically learn features from the data let to the development of the autoencoder. The autoencoder is a type of unsupervised Neural Network, being part of representation learning methods, automatically learning from input data, reducing the dimensionality of the features and recreate the original dataset [64]. Autoencoders can be compared with the Principal Component Analysis, another representation learning method, in which the latter transforms multi-dimensional data into linear representations, while Autoencoders can generate non-linear representations, having the ability to catch multimodal features of the input [59].

An autoencoder is composed by two parts: the encoder and the decoder, as seen in figure 2.12, also having their customized parameters like any Neural Network. The encoder receives the input *x*, can be any type of data like images, video, or text, and maps the input into a latent encoding space, compressing the information of *x*. For example, considering an MRI slice of 256 × 256 voxels, the encoder can transform that size into a vector of 50 × 1, or other preferred size. The decoder part uses the data compressed by the encoder into the latent code and tries to reconstruct the original input *x'*, having the same size as the input. These two structures can use architectures like any other Neural Network model, from fully connected layers to convolutional and pooling layers, so by increasing their complexity, autoencoders can learn more complex features from the original data [64]. Even though autoencoder learning is based on backpropagation, commonly used in supervised training, they are considered

unsupervised DNNs because the input is restored after the decoder part, instead of using different sets of target values [59].

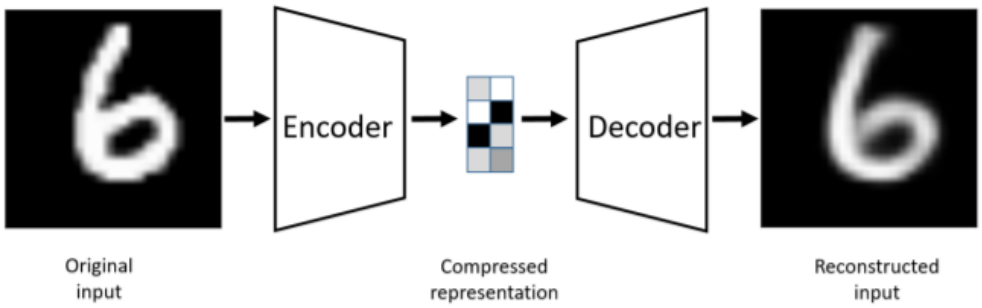


Figure 2.12: The structure of a simple autoencoder [65].

The main goal of an autoencoder is to reconstruct the input as precisely as possible only with the most important features, being this achieved by using the reconstruction loss during the training of the model, which typically is the MSE between the input and output, as demonstrated in equation 2.23, penalizing the model when output x' differs from the input x. The real purpose of using the autoencoder is not to generate a perfect copy of the input, it is in fact expected that the latent encoding space, where the compression of data is done, originate less redundant features, allowing to reduce the dimensionality of the input data, which is very important to avoid computational time and, more importantly, the overfitting problem of ML and DL models [64].

$$Loss\ function = \|x - x'\|^2 \quad (2.23)$$

The simplest way to discover valuable features from the input data is by using the so called undercomplete autoencoder. This type of autoencoder restraints the architecture, with the latent encoding space having less neurons than the input, leading to the encoder structure to compress the input data, performing data dimensionality reduction [65]. There are other variations of autoencoders created, like the denoising autoencoder, the sparse autoencoder, and the adversarial autoencoder, which are not approached in this work. For a detailed description of these methods, the reader is referred to [64,65].

Autoencoders have been also widely used to pretrain the Neural Networks, allowing a better tuning of the parameters that consequently lead to an improvement of the Neural Networks training process. Another use of autoencoders in Neural Networks is as a regularization technique for a classification network, where the network is connected to the encoder and decoder [65]. Furthermore, the ability that this technique has to automatically extract useful features from the input data and reduce feature dimensionality is extremely important in brain disorders research, due to the high dimensionality, complexity and sometimes small dataset sizes involving neuroimaging data [64].

2.3.4 – Model Evaluation

After the model classification is finished, it is extremely important to evaluate how precise and reliable the model predictions are regarding the data used. First, the dataset is partitioned into different sets in order to study model's performance, with two options for that partition: the two-way split, where the dataset divided into training and test set, or the three-way split, used to estimate the performance of the model when tunning hyperparameters, splitting into training, validation, and test set. Training set is a specific amount of data from the whole dataset from used to train the DL model, helping to fit the parameters to the model and optimize the classification. The training set consists of input data paired with the correct corresponding output, known as the target or label. Validation set is a predetermined

sample of data from the whole dataset, different from training and test set, used to perform an unbiased evaluation of the trained model while tunning the model's parameters, also known as parameter tuning. Test set is a predetermined sample of data from the whole dataset, which is different from the data used in training and validation sets the previous sets, only used to evaluate the model completely trained when tested on unseen data [66,67]. The reason why using two-way split when tunning model hyperparameters is not recommended, is because these splits reuse the test set multiple times, introducing a bias, and influencing model performance, resulting in excessively optimistic estimates, whereas three-way split uses the test set exclusively for model evaluation, avoiding the previous problem [68].

The most used evaluation measures to quantitatively assess the performance of a model classification in binary or multi-class classification problems is accuracy, for being easy to use and to understand by human. Accuracy is calculated through equation 2.24, by using the number of positive and negative correctly classified instances, TP and TN, respectively, and the number of positive and negative misclassified instances, FP and FN, respectively. The values used to calculate the accuracy can be explained through the confusion matrix shown in table 1, where it relates the labels predicted against the actual labels of the data, also allowing for other evaluation measures to be applied and provide more information about the general quality of the model classification [69].

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)} \quad (2.24)$$

Despite the wide use of accuracy, there are cases when might not be the best metric to evaluate the quality of the model. Unfortunately, in the majority of the datasets, the targets or labels present in the data are not evenly distributed, creating a very common problem known as class imbalance. Let's consider that a dataset has 1000 samples, with 995 of those being negative samples and 5 as positive. If the model classifies all the samples as negative, accuracy will be 99.5%, despite the classifier being unable to catch the positive samples. Applying accuracy as the main evaluation metric of the models in imbalanced data can origin misleading conclusions, since there is a high probability that the results are biased towards the class with greater presence in the dataset, achieving a misleading higher accuracy [70].

Table 2.1: Confusion matrix for a binary classification.

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Two metrics also used to measure the performance of the model are the recall and the precision, evaluating the efficacy of a model to classify each class in the binary classification problem. Recall, also referred as sensitivity or true positive rate (TPR), measures the number of positive instances divided by the number of all relevant instances, instance that should've been classified as positive. Precision on the hand, also referred as positive predictive value, measures the number of positive instances that the model classified correctly from the number of positive instances predicted by the model, with both measures mathematically described in equations 2.25 and 2.26, respectively [66,70].

$$Recall = \frac{TP}{TP + FN} \quad (2.25)$$

$$Precision = \frac{TP}{TP + FN} \quad (2.26)$$

An alternative metric proposed was the F-score, which is calculated by using the weighted average of precision and recall, relying the weights on a constant β, as shown in equation 2.27, controlling the weights trade-off between precision and recall. In the majority of cases $\beta = 1$, making F-score more known as F_1-score, with the metric being defined through the harmonic mean between recall and precision values, as seen in equation 2.28 [67]. F_1-score ranges from 0 to +1, where the maximum value indicates an excellent performance by the model, allowing to acknowledge how precise and robust the model can be, classifying the instances correctly and with the fewest number of misses [70].

$$F - score = (1 + \beta^2) \times \frac{Precision * Recall}{\beta^2(Precision + Recall)} \quad (2.27)$$

$$F_1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (2.28)$$

In current days, a very common evaluation metric used in binary classification problems is the Receiver Operating Characteristics (ROC), given by the relation between the sensitivity, or TPR, and the false positive rate (FPR), also known as the opposite of specificity, corresponds to the portion of negative instances that are wrongly classified as positive, considering all the negative instances of the data, as defined in equation 2.29. This metric has become widely used to assess the classification performance of models because it does not suffer the limitations faced by accuracy, namely the class imbalance problem. The main reason why ROC is an adequate evaluation metric, even in class imbalances, is that it considers the relationship between two distinct metrics, TPR and FPR, in a single metric, while other model evaluation measures like precision and recall focus only on the performance of each class, with this performance being evaluated in two distinct measures [71].

$$FPR = \frac{FP}{FP + TN} \quad (2.29)$$

The ROC is represented through a two-dimensional graph that plots the TPR in function of the FPR, with the points for each metric being then joined to create a curve, indicating the ROC curve of the model, as seen in figure 2.13. The ROC curve offers a visual tool to evaluate the capability of the model to correctly identify the positive and negative instances that were incorrectly classified [69]. The higher performance of model classification is related to the curve proximity to the top-left corner of the ROC space, where the FPR is small and the TPR is larger [71]. The information about the model classification performance in the ROC curve can be quantified by calculating the area beneath the ROC curve, a metric known as area under the ROC curve (AUC or AUROC) [70]. The AUC ranges from 0 to +1, where the greater the value, the better is the classification performance of the model in distinguish between the positive and negative classes.

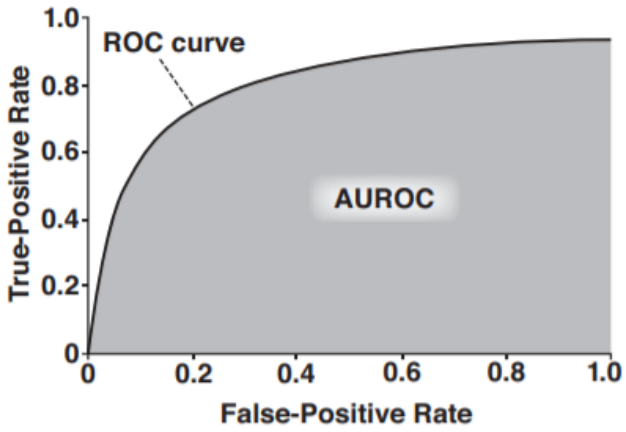


Figure 2.13: Example of the Receiver Operating Characteristics graph and the area under the Receiver Operating Characteristics curve [72].

Most of times, especially when dealing with medical data, the quantity of data is limited, which can make DL models less effective as they exhibit the best performances when trained in huge amount of data. Due to the problem mentioned above, the available data has to be wisely reused, in order to estimate the performance of the models in the most trustworthy way possible, avoiding the overfitting or underfitting generalization problems [71].

The most common data resampling technique for model evaluation and selection is cross-validation, where the main concept is that each instance of data present in dataset can be tested by the model. One type of cross-validation is the *k*-fold cross-validation, which randomly splits the same into *k* parts, with one *k* part to be used as validation set and the other $k - 1$ parts are stacked into a single training set, as shown in figure 2.16, iterating over the dataset *k* number of times [73]. Is important to point out that each data instance is only used once as validation set along the different *k* iterations, guaranteeing no overlaps. The *k*-fold cross-validation will lead to *k* different models, where each one is fitted with distinct samples of the dataset, with the model performance of each evaluation metric being the arithmetic mean of the *k* models performances for that respective metric, as demonstrated in figure 2.14 by using 5 folds [68]. When the dataset is imbalanced, *k*-fold cross-validation can create problems, since the minority class may not even be represented in one of the *k* folds, leading to misleading overly optimistic performances. To avoid this problem, the stratified *k*-fold cross-validation is used, ensuring that the classes proportions are the same in both training and validation folds [71,73].

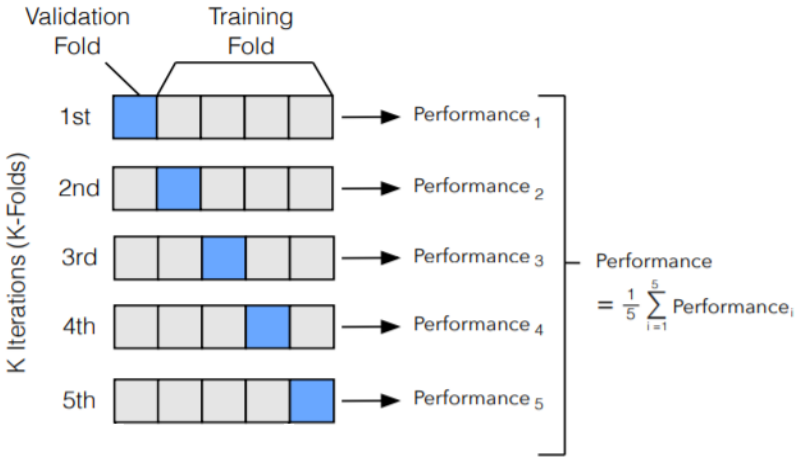


Figure 2.14: Illustration of k-folds cross-validation process for 5 folds. Adapted from [68].

2.3.5 – Black-box Problem

As mentioned previously, the importance of AI and ML techniques, like DNNs algorithms, are becoming more and more important in medical settings, mainly in diseases diagnosis and prognosis. These DNNs can provide exceptional classification accuracies in numerous complex medical tasks, from image to signals analysis, but despite that performance, these models are not highly transparent [74].

The lack of transparency from DNNs models arises from the nature of the respective algorithms, where the explainability is sacrificed for prediction accuracy, with these models learning important features by themselves instead of being chosen by the developer. With the increasing of layers and complex connections across the network, which leads to highly nonlinear associations between inputs and outputs, turn out to be very difficult to understand among users [75]. This problem is seen as a black-box, exemplified in figure 2.15, where the model receives an input data and provides the decision, without knowing the internal inference processes which led to the use of certain input information to accomplish the outcome [76].

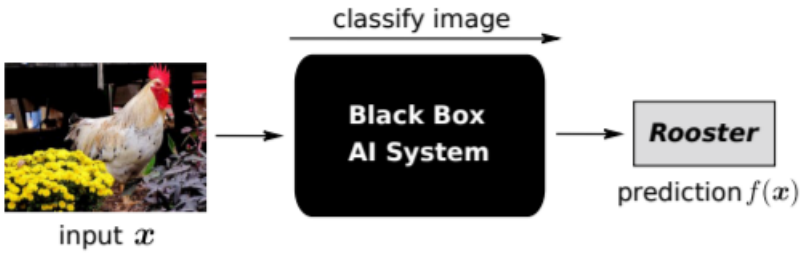


Figure 2.15: Artificial Intelligence systems prediction scheme, where nothing is known about what led to the prediction f(x) of an input x [76].

The black-box problem does not influence the quality of the DNNs models directly, rather it creates problems when it comes to evaluating what information the model is using to achieve those results. The problem with these models is that their architecture is projected to learn from the training data supplied, which are not always perfect, having a probability of presence of biased data, with the model consequently learn that faulty characteristics as well. An example of that problem is a classifier created to distinguish enemy from friendly tanks, which presented high accuracy results but in practice did not execute the proposed task. Instead of classifying enemy from friendly tanks, it was actually a good weather classifier, because the images for the training set of enemy tanks were from cloudy days and the images from the friendly ones were taken in sunny days. Another example is the dog or wolf classifier, where it revealed to be a snow classifier due to the influence of snow background in dogs and wolfs images from the training set [77].

So, with the purpose of avoid this black-box problem among DNNs and other AI models, emerged a new field called XAI, which is an AI with the objective of provide an easier understanding, analysis and most of all an explanation, for both experts and mere users, about why the model made that outcome on that problem [74]. With the increasingly development of these techniques to be used in medical environment, it is extremely important to provide the most transparency possible to all involved, discussed in [78] and [79], mainly the clinicians, as they need to have a strong foundation about the decision-making process occurred during the automated diagnosis, for relying on that relevant clinical information and explain their decisions [76].

2.3.5.1 –Explainable AI Methods

After the model perform the classification and the output predictions are known, the XAI methods are applied to explain which input features contribute positively or negatively to the prediction obtained. It is important to highlight the differences between the concepts of explainability and interpretability, often misunderstood. Interpretability consists in mapping something, for example a predicted class, so that it is visually perceptible to the human being, while explainability can include the interpretability and is the collection of features that have contributed to a certain classification prediction, in terms of relevance values, which can be observed later through a heatmap overlaid on the input [80].

The challenges faced in the complexity of analyzing the DNNs led to the expansion of the field of XAI, where different approaches were proposed to offer a diverse set of options to explain the classification predictions obtained. There are two types of model's explanation: the *ante-hoc* or intrinsically interpretable explanation, responsible for giving explanations from the beginning of the model and allowing to assess how correct a neuron in the network is about his prediction; and the *post-hoc* explanation, which evaluates the explainability of a model from its outcome, revealing the input data responsible for the final decision [77].

In explanations of DNNs, there are three major groups of methods: visualization, model distillation and intrinsic methods. Visualization methods provide an explanation about the output of a DNN by visually highlighting characteristics of an important input feature. Model distillation methods use a white-box model that is created to simulate the input-output relationship of the DNN used, allowing to identify the input features that influence the outputs. Finally, the intrinsic methods are DNNs that have been specifically designed to achieve an explanation along with the output result [81]. In the scope of this project, visualization methods will be addressed, which comprise two types: backpropagation-based and perturbation-based. The main focus will be on backpropagation-based methods, where the relevance of input features is evaluated based on the volume of gradient passed across the network layers during training, namely in LRP method.

One of the most recent XAI technique developed to explain the model's classification predictions is the LRP, a backpropagation-based method implemented based on the decomposition principle [77]. One of the advantages of LRP in comparison to other backpropagation-based or perturbation-based methods relies on the fact that the other methods measure the difference in response in output's prediction when the input features are changed, while LRP measures the relevance strength between the input feature and the specific output prediction, without any change in network inputs [81].

Given an input x and the computation of the prediction $f(x)$ in the output layer, the principle ensures that the prediction is fully redistributed backwards through all the layers of the network until the input variables are reached, with a relevance score R_j being established for each input variable, for example, for each pixel in an image, as observed in the scheme of figure 2.16 [76]. The relevance scores allow to know how much an input variable impacts the prediction, either positively or negatively, which can be then displayed through a heatmap of the original input data.

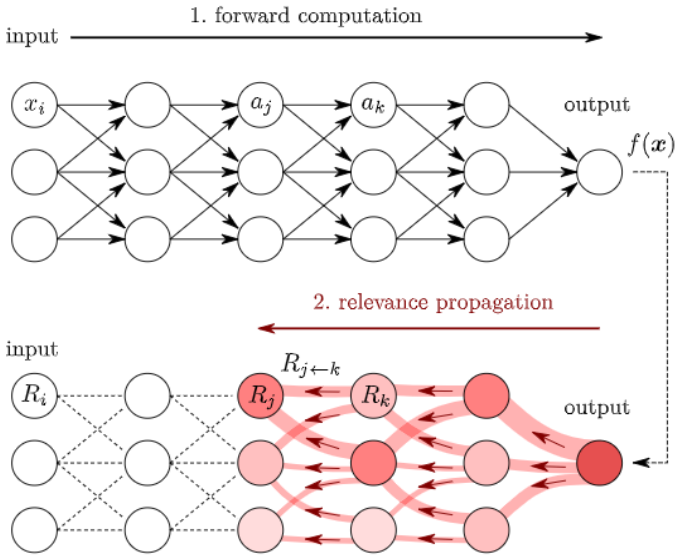


Figure 2.16: Layer-wise Relevance Propagation application in a Neural Network. Adapted from [80].

LRP can be implemented through different rules, but the most basic rule of LRP is described by equation 2.30, where the relevance is redistributed evenly from a layer l+1 to layer l, with a_j being the neuron activation at layer l, R_j as the relevance score from the neurons of the previous layer and w_{jk} is the weight of the connection between neuron j and neuron k in the layer above. The rules of LRP have a property of relevance conservation, which means that the relevance score R_j of a prediction $f(x)$ is the same for every layer backpropagated in the model [76]. The denominator of each rule is responsible for the enforcement of the relevance conservation property [82].

$$R_j = \sum_k \frac{a_j . w_{jk}}{\sum_{0,j} a_j . w_{jk}} R_k \quad (2.30)$$

Later, other LRP rules were proposed, since the gradients of DNNs, responsible for the correct update in the direction and quantity of the network weights, are usually noisy. Starting by the LRP_{ε} (epsilon) rule, shown in equation 2.31 and an improvement of equation 2.30, with the main difference being the addition of a small constant ε to the denominator to avoid divisions by zero. The constant ε absorbs some relevance when there are inconsistent or weak contributions to the activation of a certain neuron, reducing the noise in the explanations and consequently in the observable heatmap [82,83].

$$R_j = \sum_k \frac{a_j . w_{jk}}{\varepsilon + \sum_{0,j} a_j . w_{jk}} R_k \quad (2.31)$$

Following the LRP_{ε} rule, another enhancement was made towards the latter, originating the LRP_{γ} (gamma) rule, given by equation 2.32, which allows to highlight the effect of positive contributions over the negative contributions from the input variables in the explanation. The parameter γ is responsible for manage the strength of influence from positive contributions, serving in more stable explanations [82]. Originally, the idea of considering positive and negative contributions separately was created with another rule, called $\text{LRP}_{\alpha\beta}$ (alpha-beta) rule, represented in equation 2.33, where the α parameter controls the strength of positive contributions and the parameter β controls the strength of negative contributions [82,83].

$$R_j = \sum_k \frac{a_j.(w_{jk} + w_{jk}^+)}{\sum_{0,j} a_j.(w_{jk} + w_{jk}^+)} R_k \quad (2.32)$$

$$R_j = \sum_k \left(\alpha . \frac{(a_j.w_{jk}^+)}{\sum_{0,j}(a_j.w_{jk}^+)} + \beta . \frac{(a_j.w_{jk}^-)}{\sum_{0,j}(a_j.w_{jk}^-)} \right) \quad (2.33)$$

In figure 2.17, is possible to compare the performance of the different LRP rules, mainly the basic LRP rule, the LRP_ε rule and LRP_γ rule. When the basic LRP rule is used, for explaining the class "castle" in the input image, it's not able to focus on that class, picking too many artifacts from the image, while LRP_γ rule is easier to understand but considers unrelated classes. LRP_ε rule is able to provide a faithful explanation of the positively contributing input features of the desired class "castle", being also able to highlight the input features that are not related with the class "castle".

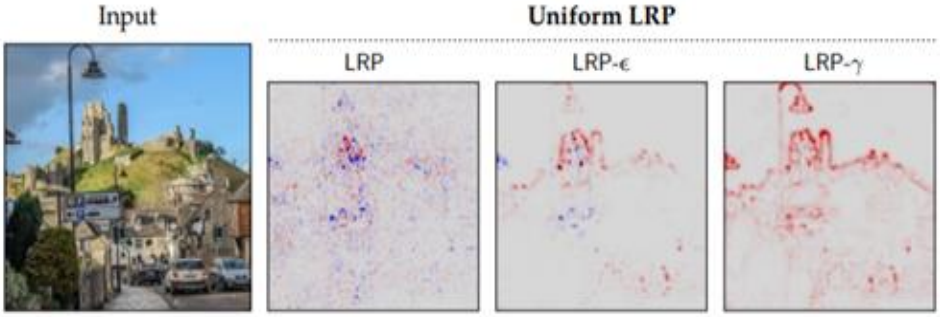


Figure 2.17: Comparison of Layer-wise Relevance Propagation rules in the explanation of input image by considering the output class "castle". Adapted from [82].

3 – State-of-the-Art

Before starting the study, it is important to recap the main objectives of this dissertation, in order to carry out a complete research on what has been explored in previous years and the current state-of-the-art in the field of FC. The main goal of this dissertation is centered on the study of the effect of the application of multiple statistical metrics to calculate the FC and create the FC matrices, beside the traditionally used correlation metric, which will then feed the DL models as inputs, in order to classify between healthy and diseased subjects of neurological and neuropsychiatric disorders. Together with this, it is also intended to study the effect of a FC multi-metric, created through the combination of these FC matrices calculated from the different statistical metrics, in the classification of healthy and diseased subjects, using the same previous DL models.

The final goal of this dissertation is to incorporate an XAI technique to the DL models used, more specifically the LRP technique, in order to provide an explanation of the internal processes performed by the models to achieve a given prediction, overcoming the black-box problem of these models, using the previously calculated FC matrices as input, revealing which regions of the brain are most important for distinguishing between healthy and diseased subjects.

3.1 – Use of Functional Connectivity Metrics

In the vast majority of FC studies, like Tang et al (2012), Dos Santos Siqueira et al (2014), Yu et al (2017) and Heinsfeld et al (2018), the analysis of connectivity in brain disorders is traditionally performed by using the FC matrices generated from the computation of correlation-based metrics between the different brain regions BOLD signal time-series from different subjects [8]. Despite the usual practice of correlation in FC studies, it has some drawbacks. Studies have shown that there are non-linearities inherent to resting-state acquisition of BOLD signals, mainly from hemodynamic origin, affecting the timing and amplitude of the BOLD signal measured, influencing the relationships between time-series and playing a significant role in connectivity analysis, more importantly when a pathology is present [84,85]. For example, if the correlation value between two signals is low, it may be wrong to assume that there is no dependence between them, where it can simply be the case that there is no linear dependence.

This limitation paved the way for the need to characterize FC matrices using non-linear metrics, where mutual information, whose definition is explained in chapter 2.2.1, is one of those metrics capable of consider non-linearities present in BOLD signals, tackling the drawback left by correlation family metrics. Zhang et al [84] aimed to compare the use of mutual information and Pearson's correlation coefficient to compute FC networks, between all brain regions BOLD signal time-series from the rs-fMRI and test their capability to predict the intelligence quotient level of epileptic patients. The results revealed that FC based on mutual information outperformed the use of Pearson's correlation in terms of accuracy prediction, showing that the FC defined from mutual information is able to capture better features of the functional brain network than correlation. Gao et al [86] in their study, applied the mutual information between all the pairs of rs-fMRI BOLD signals to construct the FC network, in order to explain functional alterations in brain networks between healthy and schizophrenia (SZ) diseased subjects, achieving promising findings regarding potential SZ diagnosis biomarkers. Along with mutual information, h^2 is another statistical metric capable of considering non-linearities present in signals, although it has not been used in FC studies to the date of this dissertation.

Another limitation found in correlation metrics use is that it might occur by chance or due to a common cause, and not necessarily by cause-effect relationships, emerging the statement “correlation does not imply causation”. Even if one signal is dependent on another, the correlation doesn’t indicate the direction of the causality present. By capturing interactions in a bivariate sense, do not consider the effects of confounding variables, without having the ability to interpret direct connections, becoming difficult to relate neuronal signals with bidirectional interactions, the predominant interactions in the brain, not having the ability to know which signal is impacting which [28,87].

To counter the inability of correlation consider causal relationship between signals from different brain regions, transfer entropy was proposed, with more details about this statistical metric described in chapter 2.2.1. In Mäki-Marttunen et al [88], transfer entropy was used to find biomarkers that could relate large-scale disruptions of brain function to the diagnosis and prognosis of patients suffering from disorders of consciousness, along with the use of partial correlation. Their work uses brain FC matrices computed through the pairwise transfer entropy between all brain regions BOLD signals time-series from each patient, defined as ROIs by the AAL atlas. The results suggest that transfer entropy, beyond partial correlation, can detect alterations in the FC of pathological patients, contributing with an important biomarker that can account for the large-scale brain function disruption in patients. Diez et al [89] decided to study the FC intercommunication between resting-state networks in individuals with AD. The transfer entropy was applied between the time-series of the rs-fMRI BOLD signals from the brain regions defined as ROIs, across all individuals present in the dataset used, originating the FC matrices. The main finding of this study is the fact that AD individuals had a higher transfer entropy value, or information flow, between resting-state networks that the control individuals, providing information on how these networks interact with each other when a pathology is present, as this can also be applied to other brain disorders. A study performed by Kumar et al [90] aimed to validate the use of transfer entropy, as a complement to Pearson’s correlation coefficient, to predict attention performance of individuals. The FC matrix constructed was based on rs-fMRI, where it was calculated from the BOLD signals time-series across each participant’s brain regions. The results obtained showed that the information flow from transfer entropy is able to predict the attention scores from individuals, demonstrating that transfer entropy can be very valuable in the characterization of human brain functional organization and a possible helpful addition to traditional correlation analysis.

With the use of correlation, information is still lacking about signals beyond the temporal domain, such as frequency and time-frequency domains. Furthermore, correlation is sensitive to the regional hemodynamic response, which can vary from individual to individual due to vascular differences, with such variability between regions resulting in decreased interregional correlation, despite the presence of neural activity [91]. For example, two brain regions can have related neural activities but different hemodynamic responses, with the correlation between the two regions being impacted by those hemodynamic variations. In Thirion et al [92] study, was proposed a framework to detect resting-state activity networks based on FC matrices computed using coherence. The FC matrices were obtained by calculating coherence between the time-series of the BOLD signals of the pairs of brain regions, defined as ROIs, across all the individuals, with a range of frequency band defined from 0.02 to 0.1 Hz. The results of this study showed that coherence FC analysis focuses on the frequency band used and is not impacted by physiological confounds, making resting-state activities more noticeable, unlike correlation.

Correlation is a metric that captures dependences only in time domain and assumes a temporal stationarity among brain signals [93]. Studies have confirmed that brain signals are non-stationary, with FC may have more pronounced dynamics in the resting-state than in task activations, therefore, when these dynamic changes are not considered, the detection of brain functional disruptions that characterize

the disorders becomes difficult and incomplete [94]. Coherence lacks the notion of time, which makes it unable to characterize the dynamics present in resting-state BOLD signals, with researchers giving a special attention towards the study of FC in time-frequency domain.

In Chang and Glover [93], Wavelet-based coherence was used to examine the dynamic behavior in relationships between brain regions at a resting-state. The FC was computed by using the Wavelet-based coherence between each participant rs-fMRI BOLD signals time-series from the brain regions defined as ROIs. The results obtained indicate that rs-fMRI signals do have dynamic properties, and these may be hidden by stationary analysis like Pearson's correlation. Later, Yaesoubi et al [95] investigated the differences between resting-state FC between SZ and healthy patients, by incorporating frequency domain characteristics with temporal dynamics. The FC matrices were estimated through the use of Wavelet-based coherence, measuring the dependence between the time-series of each patient rs-fMRI BOLD signals. The results of this study showed promising findings, where their particular FC patterns that can distinguish between SZ and healthy patients, are only recognizable when connectivity is analyzed in both time and frequency domains, suggesting that this joint domain can be very useful in revealing differences and similarities between diseased and healthy populations. A study conducted by Al-Hiyali and colleagues [96] aimed to study the influence of dynamic FC patterns to classify individuals with autism spectrum disease (ASD). The authors used Wavelet-based coherence to compute the FC data between rs-fMRI signals from all individuals present in the database used, with this FC being afterwards represented as a scalogram image. The results obtained in this work surpass previous related studies that used Pearson's correlation coefficient to compute FC data, showing that the use of both temporal and frequency domains is able to capture information from FC data that escape the traditionally applied Pearson's correlation.

Although most studies presented apply statistical FC metrics individually to study the dependence between subject's rs-fMRI BOLD signal time-series, few studies have tried to combine together FC matrices from different statistical FC metrics and take advantage of the different sources of information that are present in each one of them. One of the first studies to combine information from diverse FC metrics was in Meszlényi et al [7], where they combined the FC matrices calculated with dynamic time warping distance and warping path length. The results demonstrated that this combination led to an improved classification performance when compared to the individual use of the respective metrics and to the correlation coefficient. More recently, a study led by Mohanty et al [97], decided to combine eight different statistical FC metrics into one single composite multi-metric, in order to perform population-based classification as well as study the relationship between FC and behavioral outcome, being those: cross-correlation, coherence, Wavelet-based coherence, mutual information, Euclidean distance, cityblock distance, dynamic time warping and earth mover's distances. The results compared the use of this multi-metric against the traditional correlation coefficient used, and the multi-metric was able to achieve better performances, being more consistent than the application of the metrics individually as well.

Based on the studies presented, three key questions still remain a hot topic of interest in FC studies: (1) Is correlation coefficient, the standard FC metric used to compute FC matrices, enough? (2) Are there other FC metrics capable of quantify FC, with the same quality or better, and provide more information than Pearson's correlation? (3) Can the use of several FC metrics together, take advantage of the bundling of different information captures, and exhibit improved performances?

3.2 – Deep Learning in Functional Connectome

A few years ago, the data generated by FC analysis began to be widely used to study and classify/predict brain disorders, namely SZ, AD and Mild Cognitive Impairment, ADHD, Epilepsy and ASD [8]. The high heterogeneity and the neural disruptions resulted from widespread connectivity networks, rather than in single brain regions, in individuals with this type of neurological and neuropsychiatric disorders, plus the amount of Neuroimaging data, can take advantage of the capabilities provided by traditional ML methods like Support Vector Machines and Linear Regression techniques, overcoming human performance in the recognition of disorders non-specific symptoms and reducing the associated error [98].

The combination between rs-fMRI with ML methods has proven to be a great promise in exposing new and important FC patterns associated to brain disorders, as well as potential biomarkers, being a valuable option in the future for their diagnosis and prediction [99]. Despite such effectiveness from ML methods in identifying associations between variables of interest, they need a significant amount of manual feature engineering, remaining a challenge due to the characteristic high dimensionality of the FC data, suffering from the overload that limits its use in applications where decisions are needed almost in real time. In addition, these methods compress the data into a feature vector. This vectorization, however, removes the spatial structure of the Connectome, a valuable source of information [100].

Recent developments in DL models showed that these methods can be very useful in complex high-dimensional datasets such as fMRI data, including FC data, since these are able to learn representations directly from the raw data, solving the problem of manual feature selection from traditional ML methods and improving the classification performance, as they naturally discover unknown patterns and can generalize better in new data [98]. Several researchers have stated their attempts to use DL algorithms, especially DNNs, on rs-fMRI in order to extract high-level FC features for the diagnosis and understanding of several neurological and neuropsychiatric disorders.

One of the first studies using DL models in classification tasks using fMRI data was from Kuang et al [101], where a Deep Belief Network was tested on rs-fMRI FC data to predict the presence or absence of ADHD, using data from the ADHD-200 Global Competition. The classification accuracy from the proposed method is better than the results obtained in the competition, giving good evidence for the use of DL in patient classification using fMRI data, namely FC data. Following this study, Kim et al [102] developed a DNN model for the classification of SZ from healthy individuals, using the FC patterns originated from rs-fMRI data. The DNN model comprises several hidden layers, with the output layer consisting in a softmax activation layer. Moreover, the authors proposed the use of a normalization technique (L_1-norm) for weight sparsity control in each hidden layers of the model. The classification performance was evaluated using a 5-fold cross-validation, comparing the proposed model with an SVM model, which had a maximum accuracy of 85.8% in the proposed method with three hidden layers plus the normalization technique, compared to 77.7% in the SVM model. In Heinsfeld et al [103] study, the authors investigated the FC patterns, measured using rs-fMRI, that could improve the identification of ASD patients from healthy individuals. They proposed a model with two stacked denoising autoencoders, used for an unsupervised pre-training of the model, being responsible for extracting lower dimensional data from the input dataset. Following the autoencoders is a multilayer perceptron, composed by two hidden layers with a softmax function in the output layer, which uses the knowledge acquired from these autoencoders to perform the classification task, containing the weights adjusted in autoencoders process. The results obtained achieved an accuracy of 70% for a 10-fold cross-validation, exceeding the traditional ML methods, such as Support Vector Machines, by 7%. A similar work was performed by Eslami et al [104], where they used a model with a simple autoencoder jointly with a

single-layer perceptron, called ASD-DiagNet, for the classification of patients with ASD from healthy controls, by using FC data from rs-fMRI. In addition to significantly reducing computational execution time, the classification performance of ASD-DiagNet outperforms other state-of-the-art methods, achieving a maximum accuracy for an imaging center of 82% and 70.1% accuracy for the whole dataset.

Despite being promising, DL models have some critical drawbacks inherent to their architecture. One of the challenges in these algorithms is the fact that they have a large set of parameters to be estimated, which can lead to overfitting if the number of training samples is low and increases the computational time and resources [8]. However, the use of CNNs can be helpful to tackle this limitation, as these convolutional networks are able to learn characteristics from a given pixel neighborhood structure, which usually is a 3×3 and 5×5 pixels, independently from the location of that pixel, allowing the weights of the network to be shared, becoming the model strengthened against overfitting by decreasing the number of weights trained [7]. Regardless of the advances made with CNNs, their application to Connectome data is still in its early days.

In Kawahara et al [105], a new CNN framework called BrainNetCNN was developed to be used specifically in Connectome data. This network breaks down the paradigm of the use of CNNs to extract spatial correlation within the data, with image shape. An edge-to-edge (E2E) layer is similar to a standard convolutional layer in a CNN, but is defined in terms of topological locality, combining the weights of edges that share nodes together. An edge-to-node (E2N) filter is equivalent to convolving the adjacency matrix with a spatial 1D convolutional row filter and adding the result to the transpose of the output from a 1D convolutional column filter. Similar to E2N layer, a node-to-graph (N2G) layer reduces the dimensionality of the input by taking a weighted combination of nodes to output a single scalar. The work of Brown and colleagues [106] was inspired by BrainNetCNN model, using FC data to distinguish between ASD and healthy individuals. The model consists in an element-wise layer as input layer of the network, with a Tanh activation function, followed by 6 feature maps in each of the E2E and E2N layers, in order to reduce the number of trainable parameters, both with Leaky ReLu activation functions and dropout regularizations. The final output layer is a single fully connected N2G layer with a softmax function, responsible for the classification prediction, ending up getting a maximum accuracy of 68.7%. In another study, Khosla et al [100] implemented a simple CNN model to use the features from FC matrices as inputs, in order to distinguish ASD patients and healthy controls. The model consisted in two convolutional layers, each with an ELU activation function, interspersed max-pooling layers, to down-sample the data, followed by two fully connected layers, with the last being the output layer using a sigmoid function, performing the final data classification. The proposed CNN was compared with an SVM and a fully-connected network, achieving a classification accuracy of 73.3% for a 10-fold cross-validation, outperforming the other methods. One of the most recent studies, Shahriman et al [107] presented a CNN model for binary classification between SZ patients and healthy controls, based on EEG FC brain network. The model used in the study is composed of two convolutional layers with a ReLU activation, followed by two max-pooling layers, acting as a feature extractor. After these layers, two fully connected layers are employed, the first uses a ReLU activation with a dropout regularization, feeding the second fully connected layer with a softmax activation function to provide the classification outcome. The CNN model used in this study was compared against an Artificial Neural Network model, achieving an accuracy of 85.81% when applied a 5-fold cross-validation, overcoming the 76.96% accuracy for the Artificial Neural Network model.

The final challenge in DL models, including CNNs, is the explainability of the classification outcomes. Since these models are treated as black-boxes, due to the use of non-linear transformations on the raw data features to map them into higher levels of abstraction, huge number of parameters to be trained and their complex architecture, it is difficult to get information about which input features are

used to support the decision on a given outcome, not providing a clear knowledge about neuroanatomical and neurofunctional changes [8, 98]. The solution for this problem is the use of XAI methods, which is critical for a future adoption in healthcare applications, since a medical diagnosis needs to be clear, understandable, and explainable to be trustworthy by physicians and patients, explaining the logic behind a certain decision. The explainability, united with the remarkable performance of DL methods, is the missing piece for its safer and trustable application in real world healthcare. The most promising practice of XAI methods used in medical imaging data is displaying a heatmap representation of the input data, indicating the importance of each voxel, or data feature, in a given classification outcome.

Over the years, several XAI methods were proposed to explain the predictions of CNNs and other DNNs, in other words, visualize what is learned by the model, beside the most commonly used methods such as extraction of activations during convolution or the visualization of networks weights. The most used XAI methods for visualization include several techniques, one of the first being the sensitivity analysis proposed by Simonyan et al [108], which works by evaluating the correlation between the uncertainty in the output of a predictor and the uncertainty present in model's inputs. Another widely used XAI method is the guided backpropagation developed by Springenberg et al [109], which computes the gradient of the score for a certain output class in relation to the input given, backpropagating only the positive values of gradient, setting the negative ones to zero, in order to obtain the input data heatmap. A different XAI method to visually explain the classification of a model is the occlusion analysis implemented by Zeiler and Fergus [110], based on the modification or omission of input features and comparing the output prediction between the original and modified input, testing how the model responds to a certain input. The aforementioned methods measure the susceptibility of the output according to input modifications, which can lead to inaccurate input features on which the DNN supports its prediction decision [81]. A powerful method to overcome this limitation is the LRP method, proposed by Bach et al [111] and whose definition was given in chapter 2.3.5.1, with its main advantage being based on the fact that considers the model's weights and output layers neuron activations, being less prone to group effects in the explanation [112].

LRP analysis has been applied to several areas, including structural MRI data as in the study by Böhle et al [112], but few studies have been conducted on the use of LRP in fMRI data for clinical disease classification, more specifically using FC data, as of the date of this dissertation. One of the studies applying LRP to FC data is the one by Yan et al [113], whose work proposes a DNN model plus the LRP method to classify SZ patients from healthy controls, based on the FC network patterns from rs-fMRI. The proposed framework, in addition to having an excellent classification performance, also allowed the identification of the most significant FC patterns among the different brain regions, by using the LRP, which would not be possible analyzing only the predictions of the DNN model. This study, together with others from different LRP applications, show the promising utility of this XAI technique in explaining DL model decisions.

Given the studies carried out to this date in the field of XAI, it is possible to conclude that there is still a lot of work to be done in the application of these techniques. This is pertinent to medical applications, since this is an area where the application of DL models is gaining enormous preponderance, working as a tool to aid clinicians in the diagnosis and prognosis of various types of diseases, including neurological and neuropsychiatric disorders. Taking this into account, some questions regarding the use of these XAI techniques along with DL models are still present, being important to improve the understanding of how these techniques operate and their future application: (1) how they analyze the internal inferences made by the models for a given input prediction? (2) can these techniques provide a clearer and reliable explanation for all parties involved in this medical procedure, from clinicians to patients?

4 – Materials and Methods

4.1 – Data Collection

In order to achieve this study objectives, the rs-fMRI data from the Autism Brain Imaging Data Exchange I dataset (ABIDE-I) was used, which belongs to the Preprocessed Connectome Project (PCP) and the International Neuroimaging Datasharing Initiative (INDI). ABIDE-I dataset is composed by 1112 subjects, with 539 of those suffering from ASD and 573 healthy controls, with this neuroimaging data being shared by 16 international institutions, including university medical centers and hospitals [114].

The data from ABIDE-I has several derivatives, where the user can choose which one to download, from preprocessed or mean preprocessed functional images, amplitude of low frequency fluctuations, Eigenvector centrality and time-series extracted from different parcellation atlases. ABIDE-I data has been preprocessed by 5 different tools, which are chosen according to the user's preferences: the Connectome Computation System (CCS), the Configurable Pipeline for the Analysis of Connectomes (CPAC), the Data Processing Assistant for Resting-State fMRI (DPARSF) and the NeuroImaging Analysis Kit (NIAK). Alongside with preprocessing tools, the noise removal strategy also has 4 different options, including band-pass filtering and global signal regression, band-pass filtering or global regression only, and neither of those strategies [115].

In this study, the preprocessed ABIDE-I rs-fMRI dataset was downloaded by using the DPARSF preprocessing pipeline, that comprised: slice timing correction, motion realignment, intensity normalization and registration of fMRI images to standard anatomical space (MNI152 space), without the application of band-pass filtering or global signal regression noise removal. The dataset was parcellated into 116 brain ROIs using the AAL atlas to extract BOLD signal time-series [115]. Downloading the rs-fMRI data through the DPARSF pre-processing pipeline resulted in a total of 879 subjects.

Another dataset used in this dissertation is the ADHD-200 Sample, also belonging to INDI, with this dataset resulting from the collaboration of 8 international institutions in order to publicly share neuroimaging data from anonymous patients diagnosed with ADHD [116]. The ADHD-200 dataset includes rs-fMRI, structural MRI, along with phenotypic information of 973 subjects, 362 of them are children and adolescents diagnosed with ADHD. It combines the three different types of ADHD (ADHD-combined, ADHD-inattentive, ADHD-hyperactive/impulsive), 585 subjects are typically developing controls and 26 subjects with unavailable diagnosis [117]. Following the same procedure as ABIDE-I dataset, ADHD-200 Sample has a preprocessed repository containing both rs-fMRI and structural fMRI, offering three different pipelines to download the preprocessed data, according to the user's preferences: ATHENA, BURNER and NIAK pipelines.

The ADHD-200 rs-fMRI data was obtained through ATHENA preprocessing pipeline, that involved the following procedures: remove the first four image volumes, slice timing correction, motion realignment, voxel-wise nuisance regression to remove variations because of physiological noise, head motion and scan drifts, with the BOLD signal time-series band-pass filtered between 0.009 Hz and 0.08 Hz. These frequencies allowed to focus only on the frequencies associated with resting-state FC, and then smoothened with a Gaussian filter. In the same way as the ABIDE-I dataset, the preprocessed

ADHD-200 was also parcellated into 116 brain ROIs using the AAL atlas, for posterior BOLD signal time-series extraction [117].

Almost all the imaging data from the original ADHD-200 Sample was included in ATHENA preprocessing pipeline download, with some subjects being excluded due to poor quality or defective values. Subjects from the Bradley Hospital/Brown University were excluded for not having a diagnosis for each subject, resulting in a total of 776 subjects obtained. It is important to highlight that in this study, the different types of ADHD were considered as only one type of ADHD.

4.1.1 – Participants

A first analysis was carried out to explore each subject's data, consisting in the observation of the values of the time-series of all subjects present in each dataset studied, being this one performed through the MATLAB toolbox called Brain Analysis using Graph Theory (BRAPH). BRAPH (version 1.0.0) is an object-oriented open-source toolbox that uses MRI, fMRI, and EEG images to perform all the steps of Graph Theory analysis. This toolbox obtains directed/undirected binary and weighted brain connectivity matrices from the image modality and atlas defined at the start, as well as perform comparisons between modular structures of the brain graph and calculate the global and local measures of the graph [118].

As a result of this assessment, some subjects from the ABIDE-I dataset presented missing values in their time-series, probably due to some errors during preprocessing or acquisition. These would lead to an incorrect FC matrix computation, due to the inability to fully relate the values of the time series of the different regions of the brain, thus conducting to the exclusion of these respective subjects. From the initial 879 subjects provided by all 16 imaging institutions, using the pipeline mentioned for this study, a total of 853 subjects from each institution remained and were used in this study, as represented in tables A.1 and A.2 of the Appendix, 393 of those diagnosed with ASD and 460 as healthy controls.

The same analysis was carried out for the ADHD-200 dataset preprocessed using the ATHENA pipeline. By using BRAPH toolbox, a few subjects from the initial 776 subjects downloaded displayed missing values in the time-series, maybe due to preprocessing or acquisition errors, which would lead to an incorrect FC matrix computation. Similarly to what was explained for the ABIDE-I dataset, these subjects were excluded from the study, resulting in a total of 768 subjects used to study this dataset, where 280 are subjects diagnosed with ADHD and 488 are typically developed controls.

4.2 – Computation of Functional Connectivity Matrices

After the extraction of preprocessed BOLD time-series from all subjects with acceptable values, Functional Connectivity matrices were calculated using another MATLAB toolbox called Multiple Connectivity Analysis (MULAN). This open-source code developed by Wang et al [119], which can be used with signals from both EEG and fMRI modalities, calculates FC matrices by applying different metrics. MULAN toolbox allows the use of 7 different families of connectivity metrics, being those from time domain like correlation, h^2, mutual information, transfer entropy and Granger causality, with coherence being from frequency and time-frequency domain, in a total of 42 possible methods.

Within the scope of this study, Granger causality methods were not considered, mainly due to the excessive computational time to perform the calculations. The final methods for each family of metrics, with the respective terminology, are presented in table 4.1, resulting in 13 FC matrices per subject analyzed.

Table 4.1: List of metrics used in the study, according to their domain and relationship between time-series.

Domain	Relationship	Metric	Methods
Time	Linear	Correlation	BCorrD (Directed bivariate correlation)
			BCorrU (Undirected bivariate correlation)
	Non-linear	h^2	BH2D (Directed bivariate h^2)
			BH2U (Undirected bivariate h^2)
		Mutual information	BMITU (Undirected bivariate mutual information)
			BMITD1 (Directed bivariate mutual information comparing individual histograms to joint histograms from 2 signals)
			BMITD2 (Directed bivariate mutual information reducing the bias of the entropy of 2 time-series)
		Transfer entropy	BTED (Directed bivariate transfer entropy)
			BTEU (Undirected bivariate transfer entropy)
Frequency	Linear	Coherence	BCohF1 (Bivariate Fourier-based coherence for min frequency)
			BCohF2 (Bivariate Fourier-based coherence for max frequency)
			BCohW1 (Bivariate Wavelet-based coherence for min frequency)
			BCohW2 (Bivariate Wavelet-based coherence for max frequency)

To perform the calculation of these Functional Connectivity metrics for both datasets used, a MATLAB (version 2020a) code was created to compute all the 13 FC matrices, using MULAN's functions developed to perform the calculations. MULAN's functions need several input parameters, some specific to certain metrics, and also *.mat* files, which are required in all metrics and have the information about each subject's 116 brain regions BOLD signal time-series. The regions of the brain involved in the 116 AAL atlas used in this study to parcellate the brain regions are presented in table 4.2, with the respective abbreviation.

Table 4.2: The 116 brain regions of the Automated Anatomical Labelling atlas template and their abbreviation.

Brain Region Name	Abbreviation	Brain Region Name	Abbreviation
Precentral gyrus	PreCG	Lingual gyrus	LING
Superior frontal gyrus	SFG	Superior occipital gyrus	SOG
Superior frontal gyrus (orbital)	ORBsup	Middle occipital gyrus	MOG
Middle frontal gyrus	MFG	Inferior occipital gyrus	IOG
Middle frontal gyrus (orbital)	ORBmid	Fusiform gyrus	FFG
Inferior frontal gyrus (opercular)	IFGoper	Postcentral gyrus	PoCG
Inferior frontal gyrus (triangular)	IFGtri	Superior parietal lobule	SPL
Inferior frontal gyrus (orbital)	IFGorb	Inferior parietal lobule	IPL
Rolandic operculum	ROL	Supramarginal gyrus	SMG
Supplementary motor area	SMA	Angular gyrus	ANG
Olfactory cortex	OLF	Precuneus	PCUN
Superior frontal gyrus (medial)	SFGmed	Paracentral lobule	PCL
Superior frontal gyrus (medial orbital)	ORBmed	Caudate	CAU
Rectus gyrus	REC	Putamen	PUT
Insula	INS	Pallidum	PAL
Anterior cingulate gyrus	ACG	Thalamus	THA
Middle cingulate gyrus	MCG	Heschl gyrus	HES
Posterior cingulate gyrus	PCG	Superior temporal gyrus	STG
Hippocampus	HIP	Temporal pole (superior)	TPOsup
Parahippocampal gyrus	PHG	Middle temporal gyrus	MTG
Amygdala	AMY	Temporal pole (medial)	TPOmed
Calcarine cortex	CAL	Inferior temporal gyrus	ITG
Cuneus	CUN	Cerebellum (3, 4_5, 6, 7b, 8, 9, 10)	Cer
Vermis (1_2, 3, 4_5, 6, 7, 8, 9, 10)	Vms	Cerebellum Crus (1, 2)	CerCrus

As mentioned above, each family of metrics has input parameters, which are very important for the consistency and accuracy of FC matrices calculation. In correlation and h^2 metrics, the parameter needed is the *model order*, referring to the number of lagged observations in the model, with h^2 having also the parameter *bins*. The mutual information has as parameters the *max delay* or *max lag*, which refers to the signaling time delays, and the *bins* parameter. Transfer entropy metric has the parameter *max lag*, that considers the signaling time delays, being the same as *max delay* used in mutual information. In coherence metric, the relevant parameters are *freqs*, indicating the desired range of frequencies to calculate the connectivity matrix with coherence, and the *fs*, which corresponds to the sampling frequency [119]. The sampling frequency *fs* is defined based on the repetition time (TR) from each international institution rs-fMRI scan, which can be found in [120] and the values are shown in tables A.3 and A.4 of the Appendix[1], for each dataset. The values chosen for the *freqs* parameter consider the evidence mentioned in 2.2.1, due to the fact that low frequency fluctuations are the basis of the rs-fMRI, and on which the data of the BOLD time-series signals under study are based [3,11]. The value of each of those parameters mentioned are presented in table 4.3, with those being chosen considering the default values provided by MULAN's authors and the best representations obtained for the data used in the study [119,121].

Table 4.3: Parameters and their values for every metric used in the study.

Metric	Parameter	Value
Correlation	model order	2
h^2	model order	2
	bins	2
Mutual Information	max delay/max lag	10
	bins	10
Transfer entropy	max delay/max lag	10
Coherence	freqs	min: 0.01 Hz max: 0.08 Hz
	fs	$\frac{1}{TR^1}$

4.3 – Automatic Classification

4.3.1 – Individual Functional Connectivity Metrics Classification

The first objective of this dissertation is to study how the distinct statistical metrics applied to compute FC data, namely FC matrices from the rs-fMRI data acquired from ABIDE-I and ADHD-200 datasets, perform in order to classify brain disorders by using DL models, evaluating the impact of these when fed with Connectome information originated from FC matrices. Alterations in brain FC have the potential to provide biomarkers to classify or predict brain disorders, detecting abnormalities that cannot be found in other imaging modalities, even when there are no significant structural changes in the brain. As described in chapter 2.2.2, the functional network is composed by ROIs, in this case are parcellated into 116 brain regions, according to the atlas used, with the FC matrices being calculated using statistical

metrics between the time-series of the 116 ROIs, creating a 116×116 array with information on the relationships between the different brain regions. One of the most used methods is to use FC matrices in classification problems as input features and feed these matrices directly into the DL model. The DL models implemented to perform the automatic classification were developed in Python (version 3.6.13), using Keras (version 2.1.6) and Tensorflow (version 1.11.0) from Graphics Processing Unit (GPU) as backend. It is important to run the DL models using local GPU instead of Central Processing Unit (CPU), since the training of the model is faster using GPU.

The first model implemented was inspired on the Connectome-Convolutional Neural Network (ConectomeCNN) model proposed by [7], being developed by researcher Antonio Cano Montes in collaboration with Instituto de Biofísica e Engenharia Biomédica. The ConnectomeCNN model starts by receiving as input the FC matrices computed by MULAN's toolbox from each dataset used, with these matrices being treated as images, coming with an input size of $[N_{subjects} \times 116 \times 116 \times N_{metrics}]$, where $N_{subjects}$ is the number of subjects from the dataset and $N_{metrics}$ the number of statistical metrics used.

The ConnectomeCNN model, shown in figure 4.1, is composed by two convolutional layers (Conv), where the first layer and second layers have a specified number of filters, with the number of filters for the second layer being twice the value of the first layer, extracting features from the input data in each convolutional layer. Usually, CNNs use squared kernel filters with a stride (3×3 and 5×5 filters), which moves by column to perform the convolution operation, resulting in a single value, allowing the same filter to be multiplied by the input data multiple times and at different locations of the input. This is very important in image classification, since important information is present in square neighborhoods of pixels because the pattern could occur both horizontally or vertically, and by using this squared filter both patterns can be extracted. However, in FC the local neighborhood is not the same as traditional images, with spatial information being useless, not obtaining any further information about the input features. This led the authors to propose a novel convolution in those two layers. In the first convolutional layer, the convolution is applied line-by-line in the input data, with a convolution filter of $[1 \times N_{ROIs}]$ size, while in the second convolution layer used the operation is applied column-by-column, with a convolution filter of $[N_{ROIs} \times 1]$ size, with both convolutional layers using a Scaled Exponential Linear Units (SELU) activation function in the end. Following the convolutional layers, feature extractors of the model, two fully connected layers (FCL) emerge, with the first being a fully connected hidden layer, with a defined number of neurons, that receives as input the features generated from the second convolutional layer, using a SELU activation function. The first fully connected hidden layer feeds the model output layer, a fully connected layer composed by two output neurons that correspond to the two classes present in data, healthy or diseased subjects, along with the application of a softmax function at this layer to calculate the probabilities of each instance belonging to a certain class.

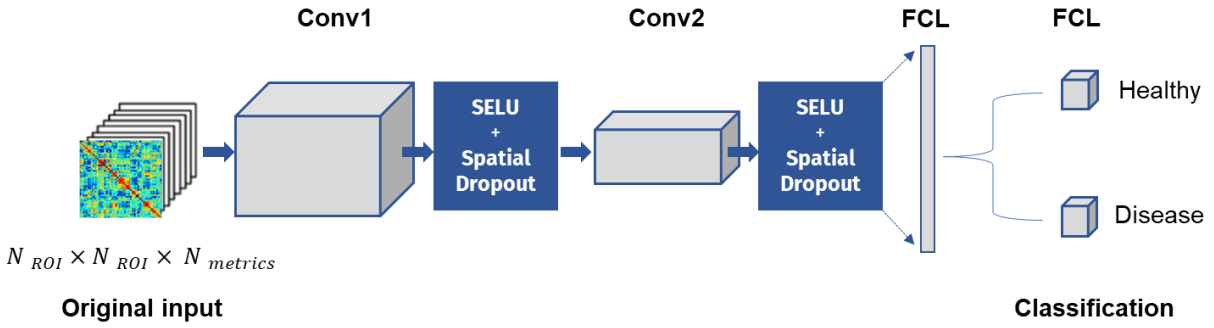


Figure 4.1: Architecture of the ConnectomeCNN model.

After each convolution layer and between the fully connected layers, dropout layers were added to conceive robustness and prevent adaptations of the model relatively to the training data used. A standard dropout layer was added after the second convolution layer and after the first fully connected hidden layer, but a different dropout technique was used in this project, namely after the first convolution layer, which in the scope of this work fits better with the idea of ConnectomeCNN. While standard dropout layers randomly drop network units in each iteration, spatial dropout layers drop the entire feature maps instead of random individual elements, as seen in figure 4.2. Spatial dropout works as a regularization technique that will consequently drop the ROI relationship with other ROIs and promotes independence between feature maps. In both standard and spatial dropouts, keep probabilities were the same.

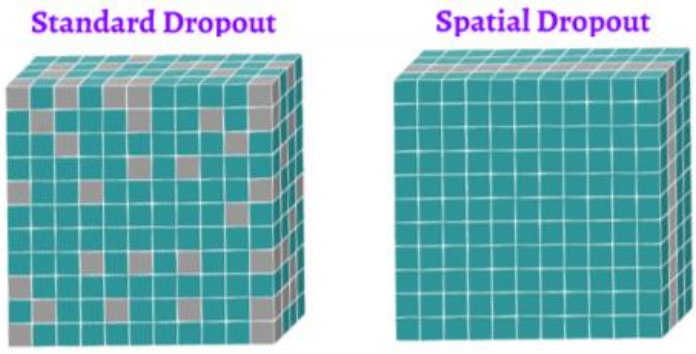


Figure 4.2:Comparison between standard and spatial dropout [122].

The second and final DL model implemented in this dissertation for the classification process was an innovative approach, developed by researcher Antonio Cano Montes who collaborated with IBEB, which was called ConnectomeCNN-Autoencoder. The purpose of this model is to combine the powerful ability of the ConnectomeCNN model to relate different brain regions with the power of autoencoders in automatic features extraction.

In ConnectomeCNN-Autoencoder model, displayed in figure 4.3, the input data size is exactly the same as the one used in ConnectomeCNN model, being $[N_{subjects} \times 116 \times 116 \times N_{metrics}]$, with the model architecture resembling the autoencoders, containing an encoding and decoding phase, as explained in chapter 2.3.3. In the encoding phase emerges the original ConnectomeCNN, where two convolutional layers are used, with the first layer applying a line-by-line convolution filter of $[1 \times N_{ROIs}]$ size to the input data, while in the second layer a column-by-column convolution filter of $[N_{ROIs} \times 1]$ size was applied, with both convolutional layers using a SELU activation function in the end. Similarly to the Connectome-CNN model, the number of filters for the second convolutional layer are twice the value of the first convolutional layer. After each convolution layer, a standard dropout is applied and provides input for the first fully connected hidden layer with the features generated, which uses a SELU activation function and feeds the two neurons output layer with a softmax function, which performs the classification. The decoder phase starts with a fully connected layer coupled to a SELU activation function, which is connected to the output layer of the encoder phase and used for the latent-space or compressed data representation, aiming to find simpler representations of the data and use them to reconstruct the original input. After the fully connected layer, three deconvolutional or transpose convolutional layers, also referred as up-sampling layers, are used to perform an inverse convolution operation to the data contained in the latent-space, taking that compressed data and transforming it into a reconstruction of the original input, keeping the latent-space data patterns. The first deconvolutional layer uses a convolutional filter of $[1 \times 1]$, while the second deconvolutional layer uses a column-by-column convolution filter of $[N_{ROIs} \times 1]$ size and the final deconvolutional layer, which is responsible

for the reconstruction of the input, uses a line-by-line convolution filter of $[1 \times N_{ROIs}]$ size. The first two deconvolutional layers use a SELU activation function, while in the last is used a Tanh activation function.

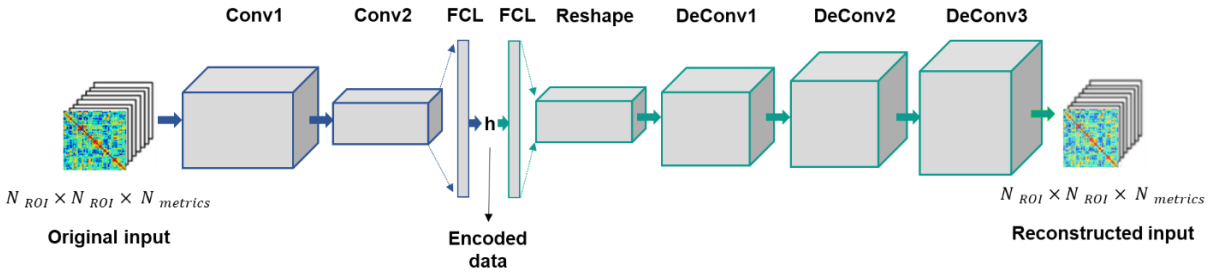


Figure 4.3: Architecture of the ConnectomeCNN-Autoencoder model.

As usual in ML and DL algorithms, to evaluate the performance of both models, the datasets were divided into two distinct sets, the training set, corresponding to 75% of the dataset used, and the test set, corresponding to 25% of the subjects present in the dataset used. Each subject was uniquely present in one of these two sets, the test set or the training set, to prevent overly optimistic classification predictions. In order to give the model the opportunity to train in different train-test splits, reducing the variance of the estimated performance when using model evaluation measures, leading to more faithful classifications, k-fold cross-validation was applied to the models used. A k=10, the most standard choice when using cross-validation as stated by Berrar in [73], splitting the data into 10 folds with different subjects in each fold train-test split. Moreover, a stratified split was applied to the k-fold cross-validation, ensuring that the classes distribution from the original dataset, healthy and diseased subjects, is maintained. Both models, ConnectomeCNN and ConnectomeCNN-Autoencoder, use a binary cross-entropy as loss function, since the output has two classes, healthy or diseased, and an Adam optimizer as learning optimization algorithm. Since the convergence of some metrics is slower than others and it's difficult to determine the proper number of epochs for the model to run, which can lead to overfitting, early stopping callback was applied to the model. In this callback, the validation loss was used as target, when this metric does not show any improvements to the validation set for 15 epochs, model training for the respective FC metric stops.

Several model evaluation measures were used to assess the performance and provide a better understanding of the models in both training and testing phases, beyond the typical accuracy. In those evaluation measures are the sensitivity or TPR, specificity, precision, F_1-score, negative positive value (NPV) and the AUC. In addition, the confusion matrix was also used to summarize the performance classification of models.

4.3.2 –Functional Connectivity Multi-Metric Classification

In addition to the objective mentioned before, it is also intended to test if the combination of those different statistical metrics used to compute the FC matrices, into one single multi-metric, can enhance the classification performance of these DL models used. The common practice in Connectome studies using FC matrices is the use of correlation family metrics to relate brain regions BOLD signal time-series from different subjects, not considering other properties inherent of these signals, as explained in chapter 3.

CNNs were designed to merge information from RGB color channels, since each channel is responsible for different characteristics of the same pixel. Transposing this ability to FC matrices, it is possible to combine different FC matrices from different statistical metrics, also keeping the respective

ROIs features information by treating the FC matrices as color channels. Even though using combined inputs can increase the number of trainable parameters of the model, as well as its complexity, combining different sources of information can enhance classification performance.

The models used to assess the classification of FC multi-metric are the same used to evaluate the impact of the individual statistical metrics, the ConnectomeCNN and ConnectomeCNN-Autoencoder models. The implementation of FC multi-metric is achieved by simply concatenate over the last dimension of the input matrix $[N_{subjects} \times 116 \times 116 \times N_{metrics}]$, as visually represented in figure 4.4, where $N_{metrics}$ will be the number of the statistical metrics used from the individual classification. The model evaluation measures used to assess the performance of both models using the FC multi-metric are the same as those used in the individual FC metrics classification, which include the traditional accuracy, sensitivity, specificity, precision, F_1-score, NPV and the AUC, along with the use of the confusion matrix.

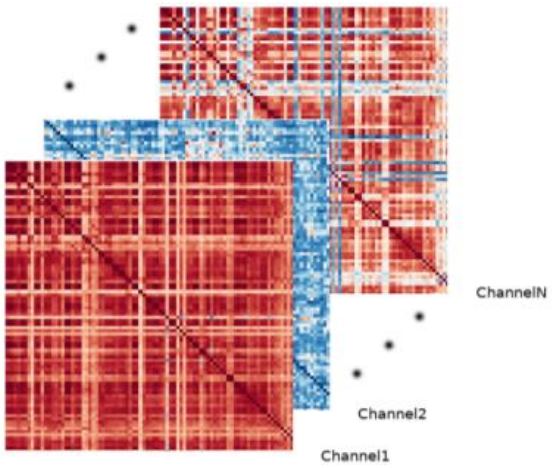


Figure 4.4: Example of concatenation between Functional Connectivity matrices computed from different statistical metrics.

4.3.3 –Optimization of Model Parameters

Once the models are developed, in order to achieve the best classification performance possible, it is extremely important to tune several model parameters, namely the number of filters or neurons present in each convolutional and fully connected layers, as well as the number of neurons to be maintained after the application of dropout layers. This is important so that these parameters are the most suitable for the FC matrix data of both datasets, used as input to the models. This set of parameters trials, shown in table 4.4, were tested for both models created, the ConnectomeCNN and ConnectomeCNN-Autoencoder models, in the two approaches studied in this dissertation, which are based on the use of each model for the evaluation of individual classification of each FC matrix computed through the different statistical metrics chosen, and on the use of the same models for the classification of a FC multi-metric.

Table 4.4: Model parameters values tested, in ConnectomeCNN and ConnectomeCNN-Autoencoder models, in order to optimize their performance for the datasets used.

	Dropout layers	**Convolutional layers**	**Fully connected layers**
Trial 1	0.65	20	42
Trial 2	0.35	20	42

Trial 3	0.65	32	64
Trial 4	0.35	32	64
Trial 5	0.35	42	80
Trial 6	0.65	42	80
Trial 7	0.35	52	92
Trial 8	0.65	52	92
Trial 9	0.35	68	102
Trial 10	0.65	68	102
Trial 11	0.35	78	114
Trial 12	0.65	78	114
Trial 13	0.35	90	130
Trial 14	0.65	90	130
Trial 15	0.35	104	162
Trial 16	0.65	104	162

From the different trials experimented, the bottom and upper range of values, namely trials 1 and 2, 15 and 16, were the limit range of values in which both models demonstrate any ability to learn the patterns of the input data. With values above and below this range, respectively, both models failed to obtain good performances, with the number of parameters of the models being too low and too complex, respectively, in order to learn significant features for the classification between the two classes, healthy or diseased subjects. Regarding dropout layers values, it was started by testing values of 0.3 and 0.7, but these neurons drop probabilities were too small and too high to cause any difference in results, leading to an exaggerated inactivation of neurons and, consequently, poor performance.

Along with the internal parameters of the models, in the case of the ConnectomeCNN and ConnectomeCNN-Autoencoder models, it is extremely important to tune the hyperparameters external to the model, namely learning rate and batch-size, with an introduction about how these models work being provided in chapter 2.3.1.1. In table 4.5 are shown the values of learning rate and batch-size tested in the ConnectomeCNN and ConnectomeCNN-Autoencoder models, in which they were tested simultaneously with changing the model's internal parameters, such as the dropout, convolutional and fully connected layers. Several batch-size values were used in the training of the created models, along with different values of learning rates, and for the data used as input, larger batch-size values led to better learning results by the models, allowing for better discrimination between healthy and diseased subjects. With the increase of the batch-size, the learning process became more consistent for the FC matrices of the different statistical metrics used.

Table 4.5: Tested learning rate and batch-size hyperparameters values for ConnectomeCNN and ConnectomeCNN-Autoencoder models optimization.

Learning Rate	Batch-size
0.001	32
0.0001	32
0.0001	64
0.001	64
0.01	64
0.001	128
0.0001	128
0.00001	128
0.001	256
0.0001	256
0.01	256
0.001	Length of training set
0.0001	Length of training set

Once the results for the various trials of parameters tested for both models and approaches used were obtained, the best configuration of parameters for each case was chosen according to its performance, being evaluated through different model evaluation measures, such as accuracy, sensitivity, specificity, precision, F_1-score, NPV and the AUC. From these model evaluation measures, it was focused primarily on accuracy results, as this is the most common measure used to evaluate model predictions, but accuracy results can be biased when the classes to be distinguished are not uniformly balanced, as mentioned previously, which is the case for ABIDE-I and ADHD-200 (more pronounced in this dataset). Together with accuracy, and to avoid biased results, AUC and confusion matrix were very important to take into account, allowing to have a better insight of the predictions for both classes, diseased or healthy subjects, preventing the influence of the larger class on the results, which can occur with accuracy. Other model evaluation measures like sensitivity, specificity, or precision, were used to perceive the model's ability to classify a particular class. The final model parameters values used in this dissertation, for each model and approaches, are shown in table 4.6. Regarding the learning rate and batch-size hyperparameters used to achieve the finest performance of the ConnectomeCNN and ConnectomeCNN-Autoencoder models, along with the use of the parameters of the models present in table 4.6, had as their final values a learning rate of 0.001 and a batch-size equal to the training set dimension used to train the models.

Table 4.6: Best configuration of parameters values for both models and respective approaches used in this study.

	ConnectomeCNN			ConnectomeCNN-Autoencoder		
Approach	Dropout layers	Convolutional layers	Fully connected layers	Dropout layers	Convolutional layers	Fully connected layers
ABIDE-I individual FC metrics	0.35	32	64	0.65	20	42
ABIDE-I FC multi-metric	0.65	20	42	0.35	20	42
ADHD-200 individual FC metrics	0.65	42	80	0.65	20	42
ADHD-200 FC multi-metric	0.65	42	80	0.35	20	42

4.4 – Explaining Model Classification

Another main objective of this dissertation is to assess which features from the FC matrices generated with the statistical metrics mentioned in chapter 4.2.1 (in this case these features are the 116×116 brain regions, since the data was parcellated into 116 brain ROIs and the FC matrix relates pairs of brain regions) are relevant for the classification of the diseased patients in ADHD-200 dataset. This is an important task, considering the fact that DL models' architecture makes them like black-boxes, being difficult to know which input features the model is really using to predict the outcome of the subjects.

To accomplish this objective, it was used the XAI technique LRP from the iNNvestigate (version 1.0.9) toolbox. The toolbox library is compatible with Python 3.6 or recent versions and is based on Keras, with a supported Keras-backend needed, which in this case is a TensorFlow backend. The iNNvestigate toolbox consists of base classes and functions that are design to implement a variety of XAI algorithms rapidly and easily along with the model created. The user only needs to adapt the algorithm already developed to the specific changes required by the toolbox, and it is up to the library to execute the desired analysis [123].

Before proceeding to the LRP analysis itself, the statistical mean between the FC matrices of all subjects for a certain statistical metric was applied, in order to study the brain regions involved in classifying subjects with ADHD as a group. The implementation process of the LRP is performed in both ConnectomeCNN and ConnectomeCNN-Autoencoder models, being described in figure 4.5, starting with the removal of softmax activation function present in the last layer of the model. Since the focus is on analyzing the model's weights before softmax activations, as it is necessary to understand how the weights of the model's neurons are considering the input data as important for the classification of ADHD, this being observed before the application of the softmax activation function. Then, the desired LRP rule is chosen to create the analyzer applied to the model trained, with this model being the

one already without the softmax activation function. As mentioned in chapter 2.3.5.1, LRP methods have a variety of rules, with each of them having a particular rule for backpropagating the relevance through the Neural Network model. After researching the iNNvestigate toolbox documentation and tests with other LRP rules, it was decided to use the LRP_{ε} rule in this study, as it is the rule with more faithful visual explanations and in terms of its relevance values. Details about how this LRP_{ε} rule works to provide model explanations can be consulted in chapter 2.3.5.1.

The last step is to use the LRP analyzer created previously to perform the relevance analysis towards the data used to test the model, test set, in order to calculate which input features from the test set are relevant for the prediction achieved. It is important to underline that the default analysis is performed examining the output of the neuron with the highest activation, existing the option of choosing which output neuron. Since in this study the classification aims to distinguish between healthy and diseased subjects, the output neuron can be one of two classes, 0 if the important features from healthy subjects want to be observed, or 1 if the features from diseased subjects are the ones to be observed. In this particular case, it is intended to observe which input features/brain regions from the FC matrices generated are considered relevant, by the model, to classify the subjects as diseased. LRP analysis works while training the model for each statistical metric, using the weights of the trained model to perform the analysis of which FC input features are considered most relevant for the desired prediction.

```
import innvestigate
import innvestigate.utils as iutils

model = create_keras_model()
model_without_softmax = iutils.keras.graph.model_wo_softmax(model)

LRP_analyzer = innvestigate.create_analyzer(LRP_rule, model_without_softmax,
                                            neuron_selection_mode="index")
OutputNeuron_index = 1

LRP_analysis = LRP_analyzer.analyze(X_test, OutputNeuron_index)
```

Figure 4.5: Example code to implement Layer-wise Relevance Propagation technique from the iNNvestigate toolbox in the models developed.

Once the LRP analysis is finished, the variable assigned will comprise the relevance values between the 116×116 brain regions for each subject present in the test set and for the respective FC statistical metric used. The whole set of statistical metrics used in individual classification are chosen to study the most relevant brain regions for the classification of diseased patients. In order to study the dataset as a group the statistical mean between the relevance values of all subjects analyzed, which are the ones present in test set, was applied, yielding a variable with 116×116 brain regions relevance values, for each statistical metric. A key step to further analyze the relevance between brain regions is the removal of the diagonal from the previous variable, as the diagonal of FC matrices relate the brain regions to themselves, having no important information on how these are related, thus ending up with a variable of 116×115 brain regions relevance values.

In addition to obtaining the relevance values between the different regions of the brain, LRP analysis allows to reconstruct the FC matrices used as input data, in the form of heatmaps. The heatmaps consist in the same dimensions of a FC matrix, but instead of relating brain regions from a statistical dependence point-of-view, it relates them accordingly with the relevance between each brain region, where a color means a strong relevance among two brain regions, while the other color means a weaker relevance among two brain regions.

5 – Results and Discussion

5.1 – Individual Functional Connectivity Metrics

The first goal of this dissertation is to evaluate how the use of FC matrices, constructed from different statistical metrics to evaluate the statistical dependences between BOLD signals time-series, in addition to the standard statistical metric correlation coefficient. These FC matrices will be used as input data for automatic subject classification using DNN models, in this case subjects with ADHD and ASD, from ADHD-200 and ABIDE-I datasets, respectively. It is important to highlight that this study's main goal was not to achieve the best possible accuracy or performance, but rather to focus on seeing whether FC matrices computed with different statistical metrics can provide valuable information to distinguish between pathological or healthy states.

For the first step, the FC matrices were computed using the procedure and parameters described in chapter 4.2.1, having a total of 13 matrices for each dataset. In figure 5.1, a FC matrix of a random subject from the ABIDE-I and ADHD-200 datasets is represented, calculated using the correlation-based BCorrU method. An example of FC matrices computed using each method applied in this study, for the same random subjects from ABIDE-I and ADHD-200 datasets used to exemplify the FC matrix of BCorrU method, is illustrated in figure A.1 of the Appendix. As seen in the figure below, the FC matrix relates the 116 brain regions, parcellated using the 116 AAL atlas, through the statistical metric chosen, leading to a range of values specific from the statistical metric used, which quantifies the statistical dependence between rs-fMRI BOLD signals from different brain regions. These FC matrices of each statistical metric computed are then applied as input data to the model, which will later classify between healthy and diseased subjects.

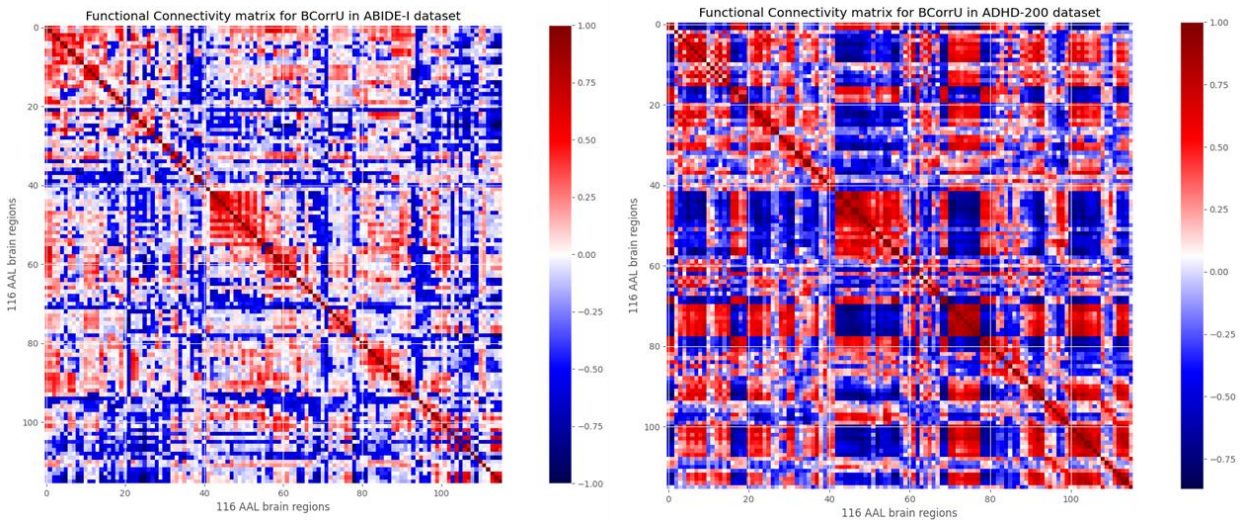


Figure 5.1: Functional Connectivity matrices examples, computed using the undirected bivariate correlation method, for a random subject from the ABIDE-I dataset (left image) and a random subject from the ADHD-200 dataset (right image).

The classification results for the ABIDE-I dataset using the ConnectomeCNN model and cross-validation are shown in table 5.1, being based on model evaluation measures such as accuracy, specificity, sensitivity, precision, AUC and the confusion matrix. It is important to refer that due to the use of cross-validation to evaluate the performance of the models, confusion matrix can contain non-integer values, since this process evaluates model's performance though the arithmetic mean between

the result from each fold. A description of how this ConnectomeCNN model is implemented and works, is given in chapter 4.3.1. As can be seen from the table, there are several FC metrics that stood out from the rest in terms of being able to correctly distinguish between the two classes, healthy or ASD subjects, namely BCorrU, BCorrD, BH2U, BH2D, BMITD1 and BMITD2. The correlation-based methods BCorrU and BCorrD, used as baseline since they're the traditional statistical metric used in FC studies, achieved an accuracy of 64% and 62%, with an AUC performance of 0.63 and 0.62, respectively. The h^2 metric, composed of the BH2U and BH2D methods, which is responsible for capturing the non-linearities of the signals, can match the performance of the correlation-based methods, obtaining both an accuracy value of 64%, plus an AUC of 0.64 and 0.63, respectively. The h^2-based method BH2U slightly outperformed the best correlation method BCorrU. An additional interesting result is the performance from the mutual information methods BMITD1 and BMITD2. These two methods achieved a classification performance similar to the correlation-based method BCorrD, ending up with 61% accuracy and 0.61 AUC, which can be justified by the sensibility of mutual information in capturing non-linear relationships between the time-series of BOLD signals from brain regions.

Interestingly, these results demonstrate that correlation is not the only FC metric capable of achieving good classification results, showing that the features present in the FC matrices generated with these methods are important to the model and are useful to discriminate the FC information between healthy subjects and those with ASD. These findings validate the limitation of using correlation-based methods, that consists in not capturing non-linearities inherent in acquired resting-state BOLD signals, having these an important role in the relationships between BOLD signals time-series, as stated in [84] and [85]. Based on the results of non-linear metrics such as mutual information based BMITD1 and BMITD2 methods, as well as h^2-based methods, these assume an important complement to correlation-based methods.

Despite the better performance of previous statistical metrics, FC matrices computed using other mutual information based methods like BMITU, plus the transfer entropy methods, were a little behind compared to previous methods. This is verified in terms of learning important characteristics of the data that would better distinguish between healthy individuals and those with ASD, achieving accuracies of 58%, 51% and 52%, respectively. Along with the lower performance of these metrics are the coherence-based methods, which are responsible to consider the time-series of the BOLD signals from brain regions in frequency and time-frequency domain. For time domain Fourier-based coherence methods BCohF1 and BCohF2, the accuracy obtained was 57% and 54% for each method, respectively, while in time-frequency domain Wavelet-based coherence methods BCohW1 and BCohW2, the accuracy was 51% for both methods. One hypothesis for the poor performances obtained by these methods is the fact that the model may have become adapted to the training data, falling into an overfitting state, with the model being unable to generalize the classification performance when confronted with data that has not been seen by the model before.

Table 5.1: Results for the classification of the ABIDE-I dataset using individual Functional Connectivity matrices and the ConnectomeCNN model.

	ConnectomeCNN					
Method	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**AUC**	**Confusion Matrix**
BCorrU	64%	69%	57%	62%	0.63	[56.4 35.3 42.6 79.7]
BCorrD	62%	64%	60%	59%	0.62	[59.1 40.9 39.9 74.1]

BCohF1	57%	61%	53%	54%	0.57	[52.5 45.2 46.5 69.8]
BCohW1	54%	55%	53%	50%	0.54	[52.6 52 46.4 63]
BCohF2	54%	61%	47%	51%	0.54	[46.6 45.2 52.4 69.8]
BCohW2	54%	56%	52%	50%	0.54	[51.3 50.3 47.7 64.7]
BH2U	64%	69%	59%	62%	0.64	[58.4 36.2 40.6 78.8]
BH2D	64%	68%	59%	61%	0.63	[58.2 37 40.8 78]
BMITU	58%	64%	52%	55%	0.58	[51.7 41.7 47.3 73.3]
BMITD1	61%	60%	61%	57%	0.61	[60.5 45.6 38.5 69.4]
BMITD2	61%	61%	62%	57%	0.61	[61.1 45.4 37.9 69.6]
BTEU	51%	56%	46%	47%	0.51	[45.6 50.5 53.4 64.5]
BTED	52%	58%	45%	48%	0.51	[44.6 48.4 54.4 66.6]

For the ADHD-200 dataset, it is even more important to consider all model evaluation measures before drawing conclusions, since this dataset has the particularity of being unbalanced. This is observed by the number of healthy and diseased subjects, with 280 subjects diagnosed with ADHD and 488 as typically developed controls, which would make the accuracy a highly biased evaluation metric to consider for assessing the performance of each method. That said, the best model evaluation metric to be considered in this dataset is arguably the AUC, which is a measure that calculates the area beneath the ROC curve, being the latter a trade-off between the sensitivity and the FPR, the opposite value of specificity, discarding overoptimistic performances in relation to the influence of a dominant class [71].

The results obtained using ConnectomeCNN model applied to ADHD-200 dataset with cross-validation, are in its majority slightly worse when compared to those obtained for the same model in the ABIDE-I dataset, confirming the difficulties in accurately classify ADHD data. As observable in table 5.2, the standard metric used in FC studies, correlation-based methods, achieved accuracy measures equal to those obtained for the ABIDE-I dataset data using the same model, with 64% for BCorrU method and 62% for BCorrD method. When the AUC is considered, it is seen that the methods based on this metric are not the best, with an AUC of 0.59 and 0.57, respectively. Similarly to the results obtained for ABIDE-I dataset, mutual information based methods and h^2-based methods showed a good ability to differentiate between typically developed controls and subjects diagnosed with ADHD. These methods matched, and even surpassed the performance of the traditional correlation-based methods. In this dataset, h^2-based methods performed slightly worse than the data in ABIDE-I dataset, with 63% accuracy and 0.60 AUC for BH2U method, and 65% accuracy and 0.60 AUC for BH2D method, demonstrating a somewhat superior ability to distinguish between healthy and ADHD diagnosed subjects compared to standard correlation-based methods. The best method for this ADHD-200 dataset is the mutual information based BMITD2 method, being able to reach an accuracy of 63% and an AUC

of 0.62, providing the most consistent classification performance among all the methods used. The BMITU and BMITD1 mutual information methods, despite having a performance inferior to BMITD2 method, both were capable of obtain AUC values of 0.59, matching the results of the best correlation-based BCorrU method. Despite having lower accuracy values, these methods indicated that they could provide a more consistent classification performance when discriminating typically developed controls and ADHD subjects. These findings demonstrate that, similarly to what was obtained in the results for the ABIDE-I dataset using the same model, despite a decreased classification performance, statistical metrics that are able to consider non-linearities present in BOLD signals are extremely important for FC information. This is confirmed by the equal or, in certain cases, better performance of FC matrices computed from h^2 and mutual information based methods compared to correlation, the statistical metric commonly used in FC studies.

Once again, likewise to what was obtained in ABIDE-I dataset results, FC matrices computed using transfer entropy based methods, were not capable of distinguish between healthy and ADHD subjects, with AUC values of 0.51 for BTEU method and 0.50 for BTED method, despite a classification accuracy of 57% and 58%, respectively. Along with these methods, Wavelet-based and Fourier-based coherence methods performed poorly in ADHD-200. Within these coherence methods, the Fourier-based BCohF1 and BCohF2 methods, achieved accuracies of 58% and 60%, respectively, but AUC of 0.52 and 0.55 confirmed the poor performance of both methods. The same was verified for Wavelet-based BCohW1 and BCohW2 methods, both having accuracies of 59%, but an inferior performance in terms of AUC, both with a value of 0.54. As stated above for the results of using these methods on the data in the ABIDE-I dataset, this poor performance by these same methods may have been due to the fact that the model has fall into an overfitting, being unable to perform well on unseen data.

It is also possible to observe in the results of Table 5.2 the influence of the imbalance between the classes of subjects in the ADHD-200 dataset, diseased and healthy subjects, in the classification performance of the ConnectomeCNN model. When looking at the confusion matrices of each FC matrix computed using the different statistical metrics, it is clear that the model used had a great tendency to better learn the features of the predominant class in the ADHD-200 dataset, which is the typically developed controls, with the opposite occurring with the smallest class, subjects diagnosed with ADHD.

Table 5.2: Results for the classification of the ADHD-200 dataset using individual Functional Connectivity matrices and the ConnectomeCNN model.

	ConnectomeCNN					
Method	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**AUC**	**Confusion Matrix**
BCorrU	64%	75%	43%	50%	0.59	[30.3 30.1 39.7 91.9]
BCorrD	62%	75%	38%	47%	0.57	[26.9 30.5 43.1 91.5]
BCohF1	58%	72%	32%	40%	0.52	[22.6 33.8 47.4 88.2]
BCohW1	59%	72%	37%	43%	0.54	[25.8 34 44.2 88]
BCohF2	60%	76%	33%	44%	0.55	[23.2 29.3 46.8 92.7]
BCohW2	59%	74%	35%	43%	0.54	[24.5 32.3

						45.5 89.7]
BH2U	63%	73%	46%	50%	0.60	[32.4 32.9 37.6 89.1]
BH2D	65%	76%	45%	52%	0.60	[31.2 28.9 38.8 93.1]
BMITU	59%	61%	56%	45%	0.59	[39.2 47.4 30.8 74.6]
BMITD1	60%	63%	55%	46%	0.59	[38.7 45 31.3 77]
BMITD2	63%	66%	58%	50%	0.62	[40.7 41.4 29.3 80.6]
BTEU	57%	71%	32%	38%	0.51	[22.2 35.5 47.8 86.5]
BTED	58%	78%	23%	37%	0.50	[16.3 27.2 53.7 94.8]

The ConnectomeCNN-Autoencoder model is an innovative approach that aims to take advantage of the ConnectomeCNN model's ability to relate different brain regions and the autoencoder's ability to work as feature extractors, trying to capture the most outstanding features of the FC data and reduce the dimensionality of these. A description of how this ConnectomeCNN-Autoencoder model is implemented and works, can be found in chapter 4.3.1.

The results obtained when using cross-validation and the ConnectomeCNN-Autoencoder model to classify the ABIDE-I dataset are shown in table 5.3. Starting by analyzing the correlation-based methods, the baseline of FC studies, the results obtained are relatively close to the results of the model ConnectomeCNN for the same data. The BCorrD method achieved the best classification performance among the two correlation methods, with an accuracy of 63% and an AUC of 0.63. Here, in the ConnectomeCNN-Autoencoder model, the h^2 metric composed of the BH2U and BH2D methods, showed the best performance in the classification task of distinguish between the two classes, healthy and ASD subjects, in terms of accuracy and AUC, in relation to all other methods used. The BH2U, as the best performing method, was able to achieve an accuracy of 64% and an AUC of 0.64, while BH2D method registered an accuracy value of 63% and AUC value of 0.63, achieving an equal performance comparatively with the best correlation-based method BCorrD. As with the use of the ConnectomeCNN model in these same ABIDE-I dataset data, the mutual information based BMITD1 and BMITD2 methods, in the ConnectomeCNN-Autoencoder model, were two of the metrics that stood out from the other methods results, in addition to those mentioned so far. BMITD1 method was able to achieve an accuracy of 61% and AUC of 0.61, while BMITD2 ended with an accuracy of 62% and an AUC of 0.62. The other mutual information method, BMITU, performed a little bit worse in comparison with the two methods mentioned above, achieving an accuracy of 59%, and with an AUC of 0.58.

These results from the ConnectomeCNN-Autoencoder model are in line with what was said above regarding the results obtained with the ConnectomeCNN model for this same dataset, indicating that there are some methods that may complement the correlation family metrics in the use of FC information to discriminate the between healthy subjects and those diagnosed with ASD. Namely methods of statistical metrics that consider non-linearities present in resting-state BOLD signals, being part of those methods the mutual information and h^2 metric. These proved to be equally beneficial and important for

the analysis of FC, overcoming the inability of correlation-based metrics to capture non-linearities in these signals, as discussed in [84] and [85].

Contrasting with the results of other methods, time domain Fourier-based coherence methods, BCohF1 and BCohF2, and time-frequency Wavelet-based coherence methods, BCohW1 and BCohW2, were one of the methods with the lowest classification performance. These had difficulties in capture features that could differentiate healthy subjects from ASD subjects, leading to accuracies of 57% and 53% for the Fourier-based coherence methods, respectively, with Wavelet-based coherence methods having both 54% accuracy. Together with this decreased performance, transfer entropy methods BTEU and BTED, were the methods with worst classification performance, with an accuracy of 50% and 52%, respectively. Again, the overfitting problem may be the cause of the model's inability to be able to better classify the dataset between healthy and ASD subjects, when using these methods.

Table 5.3: Results for the classification of the ABIDE-I dataset using individual Functional Connectivity matrices and the ConnectomeCNN-Autoencoder model.

ConnectomeCNN-Autoencoder						
Method	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**AUC**	**Confusion Matrix**
BCorrU	62%	70%	53%	61%	0.62	[52.9 34.2 46.1 80.8]
BCorrD	63%	71%	54%	62%	0.63	[53.8 33.3 45.2 81.7]
BCohF1	57%	63%	50%	54%	0.56	[49.1 42.2 49.9 72.8]
BCohW1	54%	63%	45%	51%	0.54	[44.3 42.7 54.7 72.3]
BCohF2	53%	62%	43%	50%	0.53	[42.9 43.5 56.1 71.5]
BCohW2	54%	63%	43%	50%	0.53	[42.9 42.1 56.1 72.9]
BH2U	64%	71%	56%	63%	0.64	[55.5 32.8 43.5 82.2]
BH2D	63%	68%	58%	61%	0.63	[57.4 37.1 41.6 77.9]
BMITU	59%	64%	52%	56%	0.58	[51.7 41.2 47.3 73.8]
BMITD1	61%	68%	53%	59%	0.61	[52.6 36.4 46.4 78.6]
BMITD2	62%	67%	56%	59%	0.61	[55.3 38.5 43.7 76.5]
BTEU	50%	60%	39%	46%	0.50	[38.7 45.8 60.3 69.2]
BTED	52%	58%	45%	48%	0.51	[44.8 48.7 54.2 66.3]

Regarding the application of ConnectomeCNN-Autoencoder model in ADHD-200 dataset, most performances of FC methods were slightly lower when compared to using the ConnectomeCNN model on the same data, as can be seen in cross-validation results from table 5.4. Starting by looking to the standard metric in FC studies, correlation-based methods, the best classification performance was achieved by BCorrU method, one of the highest obtained, with 63% of accuracy value but with an AUC value of 0.58. Through the table below, it is possible to observe what has been noticed in previous results, where the methods based on h^2 metric and mutual information are achieving a similar, and in some cases, better classification performances in relation to the correlation-based methods. From the two h^2-based methods, BH2U method is the one with the highest ability to distinguish typically developed controls and subjects with ADHD. This method had an accuracy of 63%, while the AUC ended up with a value of 0.58, matching the best correlation method, BCorrU. Although slightly inferior, the BH2D method managed to achieve the same accuracy as BH2U method, but with an inferior AUC of 0.57. When it comes to the mutual information methods, two of the three methods were able to reach equal accuracy values in comparison with BCorrU method and both h^2 metric methods, with BMITU surpassing all with an accuracy of 64%. Despite this higher accuracy, the AUC of the BMITU method was not the best, ending up with an AUC of 0.57, together with the BMITD2 method, both very similar to the best correlation method.

Even though the use of this ConnectomeCNN-Autoencoder model in the ADHD-200 dataset has experienced a decrease in classification performances among what have been the best individual metrics in this study, such as correlation, h^2 and mutual information based methods, it is still possible to notice the proximity between the performances of these metrics. This proximity reinforces the fact that has been observed previously regarding the use of statistical metrics with the ability to consider the non-linearities of BOLD signals, like h^2 and mutual information based methods, indicating their importance and positive complement to the correlation metrics, overcoming its limitation, as shown in [84] and [85]. As a trend in the results presented so far, here in the classification results for ADHD-200 dataset using ConnectomeCNN-Autoencoder model, all coherence-based methods showed a poor performance in terms of distinguishing healthy and ADHD subjects. Although most of these methods obtained accuracies in the order of 60%, their AUC values barely surpassed the 0.50 barrier. Together with these results for coherence metric methods, are the results for the transfer entropy methods BTEU and BTED. These two methods ended up with the worst performances among all the FC statistical metrics used with the ConnectomeCNN-Autoencoder model, even with accuracies of 56% and 58%, respectively, the AUC for both models were the two worst values, being respectively 0.49 and 0.50.

It is evident for the ConnectomeCNN-Autoencoder model, identically to what happened using ConnectomeCNN model, the difficulties in distinguishing ADHD subjects from typically developed controls. In the ConnectomeCNN-Autoencoder model, this is slightly more accentuated due to the fact that the model has in its structure an autoencoder. Since the ADHD-200 dataset is quite unbalanced in terms of the number of healthy and diseased subjects, the model's autoencoder emphasize the features from the typically developed controls, the predominant class, translating this into the low results obtained for sensitivity. What was said previously can be observed in the confusion matrices of each FC matrix used in table 5.4, where there is a noticeable capacity of the model to recognize the class with the greatest presence in the dataset, similarly to what was observed for the same dataset when using the ConnectomeCNN model for classification. In addition to what was said, the presence and negative effect of overfitting is a hypothesis that may also be associated with these less positive results.

Table 5.4: Results for the classification of the ADHD-200 dataset using individual Functional Connectivity matrices and the ConnectomeCNN-Autoencoder model.

ConnectomeCNN-Autoencoder						
Method	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**AUC**	**Confusion Matrix**
BCorrU	63%	78%	38%	50%	0.58	[26.9 27.3 43.1 94.7]
BCorrD	62%	78%	33%	47%	0.56	[23.3 26.4 46.7 95.6]
BCohF1	60%	78%	28%	42%	0.53	[19.4 27.1 50.6 94.9]
BCohW1	60%	78%	27%	42%	0.53	[19 26.5 51 95.5]
BCohF2	61%	81%	26%	43%	0.53	[18 23.7 52 98.3]
BCohW2	59%	78%	25%	39%	0.51	[17.2 26.5 52.8 95.5]
BH2U	63%	78%	37%	49%	0.58	[26 26.8 44 95.2]
BH2D	63%	79%	35%	49%	0.57	[24.5 25.1 45.5 96.9]
BMITU	64%	80%	35%	50%	0.57	[24.3 24.2 45.7 97.8]
BMITD1	63%	84%	26%	49%	0.55	[18.1 19 51.9 103]
BMITD2	63%	80%	33%	49%	0.57	[23 24.2 47 97.8]
BTEU	56%	76%	23%	35%	0.49	[15.8 29.4 54.2 92.6]
BTED	58%	78%	23%	37%	0.50	[16 27.2 54 94.8]

Although the main objective of this dissertation is to test the use of different statistical metrics to calculate FC matrices, it is important to compare the classification performance of the DL models used in this study, the ConnectomeCNN and ConnectomeCNN-Autoencoder models, with other classifiers tested with these two databases used, ABIDE-I and ADHD-200. Several authors have tested their DL models on the ABIDE-I dataset, using the data coming from the CPAC pre-processing pipeline. These models in [103] and [104] achieved accuracies of 70% and 70.1% for the entire dataset, respectively. In addition, CNNs were also applied to this dataset, namely in [100] and [106] papers, each ending with a classification accuracy of 73.3% and 68.7%, respectively. When it comes to the performance of the ConnectomeCNN and ConnectomeCNN-Autoencoder models in the ABIDE-I dataset, both had the best overall performance in the FC matrix of the BH2U method, with an accuracy of 64% and an AUC of 0.64, as can be seen in tables 5.1 and 5.3, respectively, showing to be slightly behind the performance of other DL models used recently. Regarding the ADHD-200 dataset, there are few studies that use DL models to classify the dataset as a whole, as performed in this dissertation, rather than classifying the subjects of each imaging site individually. Two studies, [124] and [125], whose classification was done

for the ADHD-200 complete dataset, following the preprocessing pipeline used in this dissertation, achieved classification accuracies of 71.3% and 70.3% with the DL models proposed, respectively. Other studies have focused only on testing their proposed DL models into classifying subjects from individual imaging sites present in the ADHD-200 dataset, so they should not be considered a good comparison. About the performance of the ConnectomeCNN and ConnectomeCNN-Autoencoder models in the ADHD-200 dataset, the best accuracy performance was achieved by the FC matrix of the BH2D method, with 65%, and by the FC matrix of the BMITU method, with 64%, respectively. When comparing the results obtained by the DL models used in this dissertation and the DL models proposed in other studies, applied to the same datasets, it is possible to conclude that the performances of the former are slightly behind other DL models.

5.2 – Functional Connectivity Multi-Metric

The second main goal of this dissertation was to test the use of the FC matrices, computed previously using different statistical metrics, combined together to create a FC multi-metric. This approach arises from the idea of combining different sources of information that are captured by each individual statistical metric, which could lead to an increased ability to distinguish between healthy and pathological subjects. To successfully handle the high-dimensionality input data that combining FC matrices from different methods would lead to, DL methods emerge as a perfectly suited option to the task. These methods, given their ability to automatically extract important features, avoid the laborious work of manual feature selection that occurs when using ML techniques. The model parameters used to study the classification performance of the multi-metric approach, for both ConnectomeCNN and ConnectomeCNN-Autoencoder models, are shown in table 4.6 in chapter 4.3.3.

The FC matrices from each statistical metric computed with ABIDE-I dataset were combined, through concatenation, and used as input to feed the ConnectomeCNN model. As observable in cross-validation results from table 5.5, the result of this combination matches the results of h^2 metric BH2U method, the best performing method in the individual classification of the different methods of each metric using the ConnectomeCNN model, as demonstrated in table 5.1. This FC multi-metric approach achieved an accuracy value of 64% and an AUC value of 0.64. An important detail of the result of the FC multi-metric approach is the fact that it was able to reproduce the sensitivity measure in relation to the same measure of the best individual method, the BH2D method. This led to a sensitivity value equal to the best value of this measure in the FC individual methods, which was 62% of the mutual information based BMITD2 method. The same reproducibility was almost verified for the precision, ending up being only 1% behind the best individual method, which were the BCorrU and BH2U methods with 62%. This result from the FC multi-metric indicates the good capacity of this approach to select the best features from all the FC matrices combined, in order to differentiate between healthy and ASD subjects, which is verified by the maintenance of the classification performance of the multi-metric approach in relation to the best method of individual classification.

In the ADHD-200 dataset, the combination of the FC matrices generated from the respective dataset data led to a performance quite similar to the best performance obtained in the individual metrics classification. As shown in table 5.5 cross-validation, the FC multi-metric approach for the ADHD-200 dataset achieved an accuracy of 64%, with the best individual method accuracy, that is from the h^2 metric BH2D method, ending up obtaining 65% of accuracy, as described in table 5.2. But, as previously mentioned, this dataset has its unbalanced classes, which means that a detailed evaluation of other evaluation measures of the model is necessary, being AUC one of these important measures to assess the unbiased performance of the model. Looking at the best AUC value from individual metrics approach, which was for mutual information BMITD2 method, and multi-metric approach AUC value,

it is possible to observe that the AUC value of the latter approach was below the classification of the best individual method, registering an AUC of 0.59 against 0.62 of the BMITD2 method. When analyzing other evaluation measures of the model, which is the case of specificity and precision, it can be observed that this multi-metric approach was able to extract the most important information from the best performing individual methods. This approach obtained a result equal to the best individual method in specificity, which was the BTED method, and a result equal to the best individual method in precision, which was the BH2U method. Despite not being able to achieve the same result regarding the best individual method in sensitivity, these results are linked with was stated in the use of the FC multi-metric approach in ConnectomeCNN model, where this approach provided good evidence of having a good ability to use the information on the best statistical metrics, making it more complete.

Table 5.5: Results for the classification of the ABIDE-I and ADHD-200 datasets using Functional Connectivity multi-metric matrix and the ConnectomeCNN model.

	ConnectomeCNN					
Method	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**AUC**	**Confusion Matrix**
ABIDE-I multi-metric	64%	66%	62%	61%	0.64	[61.5 38.6 37.5 76.4]
ADHD-200 multi-metric	64%	78%	41%	52%	0.59	[28.8 27.1 41.2 94.9]

Regarding the use of ConnectomeCNN-Autoencoder model with the FC matrices combined from ABIDE-I dataset, as demonstrated in table 5.6, the performance of multi-metric approach is almost the same as the results achieved by the ConnectomeCNN model using the same approach. Both models achieved an equal accuracy value of 64%, with the AUC for the ConnectomeCNN-Autoencoder model being once again the same as the AUC obtained for the ConnectomeCNN model. In terms of the other model evaluation measures, the multi-metric approach of the ConnectomeCNN-Autoencoder model was able to slightly overcome the specificity of the same approach when used in the ConnectomeCNN model, obtaining a specificity of 68%, while the FC multi-metric of the ConnectomeCNN model obtained only 66%. The precision values were very close between the two models used with the FC multi-metric, but in sensitivity, the ConnectomeCNN-Autoencoder model was a little below the result obtained with the ConnectomeCNN model, with the latter achieving 2% more in sensitivity value.

When comparing the FC multi-metric classification against the individual metrics classification in table 5.3, using the ConnectomeCNN-Autoencoder model, it is noticeable that this approach can achieve a similar accuracy and AUC values to the respective values obtained by the best individual method, which was the BH2U method. In relation to precision, the multi-metric approach was able to reproduce the result from the best performing individual method, achieving a precision value of 62%, the same of BCorrU method. In terms of sensitivity, the FC multi-metric managed to surpass the best individual method, which was the BH2D method, with a value of 58%, and improve the result of this evaluation measure, achieving a sensitivity of 60%. On the other hand, in specificity measure, the FC multi-metric was not capable to come close to the result obtained by the best individual method, the BCorrD and BH2U methods, which ended up with a specificity of 71%, while the specificity for the multi-metric approach was 68%.

In the ADHD-200 dataset, when the FC multi-metric approach was applied using the ConnectomeCNN-Autoencoder model, the classification performance was quite similar to the best individual metrics, as well as the FC multi-metric approach using the ConnectomeCNN model. As observable in table 5.6, despite an accuracy of 64%, model's ability to discriminate between typically developed controls and subjects diagnosed with ADHD is not so good, which can be seen by the 0.59 value of AUC. The classification results for the same approach with the ConnectomeCNN model differ from the latter only in terms of sensitivity, with better sensitivity by 3%. As noted, using the multi-metric approach in the ConnectomeCNN model, comparing the result of this FC multi-metric with the results of the individual methods in table 5.4, the result of the FC multi-metric in the ConnectomeCNN-Autoencoder model resembles the performance of the individual methods with better performance in each evaluation measure of the model. This multi-metric approach using the ConnectomeCNN-Autoencoder model was able to surpass by 1% the sensitivity from the best performing individual method, the correlation-based BCorrU. It was also able to surpass the best individual method in precision measure, achieving a precision of 52%. In terms of specificity, the FC multi-metric was a little short of the best individual method performance, which was the mutual information BMITD1 method with 84%, while the multi-metric approach achieved 80%. In line with what was shown using the FC multi-metric approach with the ConnectomeCNN model, this approach with the ConnectomeCNN-Autoencoder model for the ADHD-200 dataset was also able to consider the most relevant features of the different FC multi-metric information sources, providing a more complete classification performances.

Table 5.6: Results for the classification of the ABIDE-I and ADHD-200 datasets using Functional Connectivity multi-metric matrix and the ConnectomeCNN-Autoencoder model.

ConnectomeCNN-Autoencoder						
Method	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**AUC**	**Confusion Matrix**
ABIDE-I multi-metric	64%	68%	60%	62%	0.64	[59.3 36.7 39.7 78.3]
ADHD-200 multi-metric	64%	80%	39%	52%	0.59	[25 23.9 45 98.1]

From the results obtained in this dissertation and the two studies, up to the moment of this dissertation, which sought to combine FC matrices computed through different statistical metrics, [7] and [97], despite the less good performance of the FC multi-metric using ConnectomeCNN-Autoencoder model with ADHD-200 dataset, it is possible to see that the use of FC multi-metric can be extremely important in the classification of brain disorders using FC data. This approach has shown the ability to consider different sources of information from the various statistical metrics used, achieving a classification performance practically identical to the best method when used individually. This comparable performance may suggest a more frequent use of this type of analysis in FC studies, since it seems to complete the capacity of correlation-based metrics in other domains, not preventing the continuity of the individual use of statistical metrics at the same time, so as not to compromise the new developments and the comparison between the two approaches.

Similarly to what was done for the classification using the individual FC matrices calculated through the different statistical metrics, is important to compare the classification performance of the FC multi-metric approach of the ConnectomeCNN and ConnectomeCNN-Autoencoder models used in this

dissertation, and the DL models proposed by other studies. In the ABIDE-I dataset, the multi-metric approach achieved the same 64% of accuracy and 0.64 of AUC when used with both the ConnectomeCNN and ConnectomeCNN-Autoencoder models. Comparing the performance of these models with the DL models employed by other authors in the classification of the ABIDE-I dataset, namely in [100],[103], [104] and [106], where a classification accuracy of 73.3%, 70%, 70.1%, and 68.7% were achieved, respectively, it is possible to conclude that this approach fell short of current state-of-the-art models. The multi-metric approach, for the ADHD-200 dataset, achieved the same classification performance for the ConnectomeCNN and ConnectomeCNN-Autoencoder models, with an accuracy of 64%. The two studies whose classification was done for the complete ADHD-200 dataset, [124] and [125], similarly to what was conducted in this dissertation, achieved classification accuracies of 71.3% and 70.3% with the DL models proposed, respectively, a step ahead from the models used in this study.

5.3 – Explaining ADHD Relevant Brain Regions

The final objective of this dissertation is to explore XAI methods, mainly the LRP method, in order to explain and provide transparency to the DL models. The architecture of these models makes them work like black-boxes, not providing a justification about why the model made that certain classification. The LRP was applied to the ADHD-200 dataset using both ConnectomeCNN and ConnectomeCNN-Autoencoder models. With this method it is expected to get a deeper understanding of which FC matrices input features, the relationships between different brain regions BOLD signals time-series, are considered relevant by the model for the discrimination between healthy subjects and subjects diagnosed with ADHD. For this task, only the individual FC matrices from each statistical metric were used, in order to compare how, from the capturing different signal information from these metrics, they consider the most relevant brain regions for the classification of subjects diagnosed with ADHD.

5.3.1 – LRP analysis with ConnectomeCNN model

The application of LRP method was firstly conducted on ConnectomeCNN model while it was performing the individual FC matrices classification. It is important to mention that within this relevance values, the redundant information present in the diagonal, which relates each brain region with themselves, is already removed. In order to study the brain regions involved in ADHD pathophysiology from this dataset as a group, it was performed the statistical mean between the FC matrices of all subjects for each statistical metric used. This procedure is conducted for the FC matrices of the subjects present in the test set, as the objective is to assess the most relevant brain regions for a classification related to ADHD.

As soon as the model executes the classification for each FC matrix from the set of statistical metrics used, by using the procedure explained in chapter 4.4, the relevance values among each of the 116 brain regions from the AAL atlas are obtained. Before analyzing the relevance scores, a visual representation of the LRP analysis is conducted, reconstructing the input FC matrix, for each statistical metric, in the form of an heatmap, indicating which pixels, or pairs of brain regions, are the most relevant. This heatmap is shown in the left image of figure 5.2, being compared to the original FC matrix input in the image on the right of the same figure. This LRP analysis concerns the classification of the ADHD-200 dataset through the FC matrix computed with the BCorrU method, using the ConnectomeCNN model, where is possible to observe the most relevant pixels for an ADHD diagnosis in red, while the least relevant pixels are in blue. The LRP analysis for the remaining statistical metrics FC matrices are shown in figure A.2 of the Appendix.

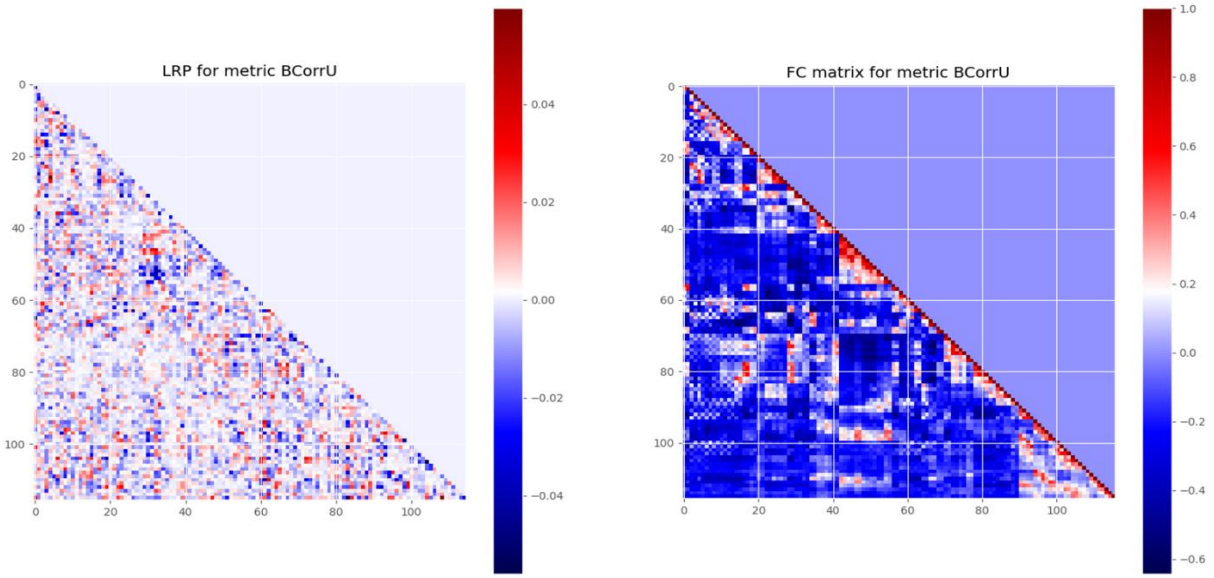


Figure 5.2: Heatmap of the Layer-wise Relevance Propagation analysis for the Functional Connectivity matrix computed with the undirected bivariate correlation method (left image) of the respective original Functional Connectivity matrix computed with the same method (right image) with the ConnectomeCNN model, using the statistical between the Functional Connectivity matrices of all subjects for this statistical method.

From the visual representation of LRP analysis using the ConnectomeCNN model emerges table 5.7, providing the ten most relevant brain regions related to an ADHD diagnosis and their respective relevance values. These values are ordered from the most relevant to the smallest among these ten regions, for each statistical metric FC matrix, considering the results of the group of subjects in the test set, presenting a more precise and quantitative evaluation of the LRP analysis.

Frontal lobe

From the results obtained in tables 5.7 and A.5, it is evident that the frontal lobe brain regions are one of the most predominant brain regions in FC matrices of each statistical metric used. These frontal lobe brain regions, composed by superior (SFG), middle (MFG), and inferior frontal gyrus (IFG), are known to be involved in several higher cognitive functions, including attention regulation. They're also involved in executive functions such as planning, organize and make decisions, coordination of voluntary movements, formation and retention of memories, conscious thoughts, and personality dysregulations, leading to changes in personality traits, with all these functions being somehow compromised when an individual is affected by ADHD [126,127]. The prefrontal cortex is the primary brain region when it comes of executive function, and dysfunctions in these regions lead to an impaired executive function, which plays a key role in the pathology of ADHD. Several studies have been reporting connectivity alterations in prefrontal cortex of individuals with ADHD, mostly discovering abnormal FC both in middle and superior frontal gyrus [128,129 ,130]. The inferior frontal gyrus is also an important region of the prefrontal cortex for ADHD pathology, since the dysfunction of this region is critical to the deficit of response inhibition, being also reported as one of the prefrontal cortex regions with connectivity alterations in ADHD individuals [130]. One of the most recent works from Riaz et al [131] showed alterations in frontal lobe in ADHD individuals, being the brain region containing the most discriminative FC activity in terms of an ADHD classification. These findings are in some extent related to a structural damage to the frontal lobe region, leading to the respective dysfunction and compromising the executive function in individuals with ADHD [128,132].

Table 5.7a: Brain regions with greater impact on the others in an ADHD-related diagnosis, when using the ConnectomeCNN model, and respective relevance values.

ConnectomeCNN										
Method	**Brain Regions**									
BCorrU	Cer 6.L	MTG.L	STG.R	THA.L	CAU.R	LING.L	CUN.L	CAL.R	CAL.L	HIP.L
Relevance Value	0.00139	0.00129	0.00107	0.00097	0.00095	0.00094	0.00072	0.0006	0.0006	0.00051
BCorrD	CerCrus 2.L	MTG.R	THA.R	SMG.L	PCG.L	IFGorb.R	IFGorb.L	ORBmed.R	SFG.R	PreCG.L
Relevance Value	0.00088	0.00086	0.00084	0.00057	0.00050	0.00045	0.00044	0.00043	0.00042	0.00034
BCohF1	Cer 8.L	Cer 4_5.L	STG.R	PUT.R	PCUN.L	IPL.R	ORBmed.R	ORBmed.L	SFGmed.R	IFGoper.R
Relevance Value	0.00123	0.00114	0.00066	0.00065	0.00064	0.00063	0.00055	0.00054	0.00054	0.00053
BCohW1	Vms 8	Cer 4_5.R	SMG.L	FFG.R	IOG.L	PCG.R	PCG.L	SFGmed.R	MFG.L	ORBmed.R
Relevance Value	0.00118	0.00104	0.00102	0.00096	0.00084	0.00081	0.0008	0.00076	0.00070	0.00067
BCohF2	Vms 6	STG.L	SMG.L	CAL.L	REC.L	ORBmid.L	SFGmed.R	IFGtri.L	IFGoper.L	ORBmid.R
Relevance Value	0.00175	0.00117	0.00109	0.00106	0.00104	0.0010	0.00097	0.00089	0.00076	0.00075
BCohW2	Cer 4_5.L	MTG.R	PCL.L	ANG.R	SMG.L	IOG.L	PCG.L	MFG.R	MFG.L	ORBmed.R
Relevance Value	0.00125	0.00119	0.00072	0.00071	0.00065	0.00063	0.00062	0.0006	0.00058	0.00057
BH2U	Vms 10	CerCrus 2.L	ANG.R	IPL.L	SOG.R	ACG.R	ACG.L	REC.R	REC.L	SMA.R
Relevance Value	0.00093	0.00079	0.00067	0.00058	0.00057	0.00055	0.0005	0.00047	0.00046	0.00045

Table 5.7b: Brain regions with greater impact on the others in an ADHD-related diagnosis, when using the ConnectomeCNN model, and respective relevance values.

ConnectomeCNN										
Method	**Brain Regions**									
BH2D	ITG.L	PHG.R	PHG.L	REC.L	ORBmid.L	SFGmed.R	OLF.L	IFGoper.L	ORBmed.L	SFG.L
Relevance Value	0.00114	0.00074	0.00064	0.00061	0.00060	0.00059	0.00059	0.00056	0.00054	0.00051
BMITU	Cer 4_5.L	CerCrus 1.L	STG.L	PCL.L	PCUN.L	SMG.L	SOG.L	PCG.L	REC.L	ORBmid.L
Relevance Value	6.75e-06	5.39e-06	5.33e-06	5.16e-06	4.34e-06	4.33e-06	4.15e-06	3.92e-06	3.80e-06	3.68e-06
BMITD1	PCL.L	ANG.L	SMG.L	PoCG.L	REC.L	ORBmid.L	ORBmed.R	ORBmed.L	PreCG.L	SOG.L
Relevance Value	0.00012	7.69e-05	4.76e-05	4.15e-05	3.70e-05	3.55e-05	3.25e-05	2.92e-05	2.83e-05	2.80e-05
BMITD2	STG.L	PCL.L	SMG.L	PoCG.L	LING.L	REC.L	ORBmid.L	ROL.R	ORBmed.L	PreCG.L
Relevance Value	0.00057	0.00047	0.00035	0.00028	0.00027	0.00027	0.00024	0.00021	0.00020	0.00019
BTEU	Vms 4_5	STG.L	HES.L	PCL.L	IPL.R	MOG.R	CAL.L	AMY.L	PCG.L	MCG.L
Relevance Value	0.00178	0.00172	0.00155	0.00148	0.00143	0.00128	0.00101	0.00094	0.00082	0.00071
BTED	Vms 9	Vms 4_5	Vms 1_2	Cer 3.L	MTG.R	TPOsup.L	FFG.L	CAL.L	AMY.L	PCG.R
Relevance Value	0.00095	0.00076	0.00074	0.00073	0.00069	0.00057	0.00054	0.00053	0.00052	0.00051

It is possible to observe that the superior frontal gyrus is the most prominent region from the prefrontal cortex, being considered relevant by almost all statistical methods used, except for the BCorrU, BH2U, BTEU and BTED methods. The middle frontal gyrus corresponds to the second region of the premotor cortex most often considered as one of the most relevant, where it was considered by the BCohF1, BCohW2, BCohF2, BH2D and all the mutual information based methods. The inferior frontal gyrus had the fewest presence in the LRP analysis of this study, being seen as relevant by the BCorrD, BCohF1, BCohF2 and BH2D methods. The location in the human brain of the frontal lobe regions considered the most relevant for a diagnosis related to ADHD, from the LRP analysis with the different FC matrices, are shown in the image on the left in figure 5.3.

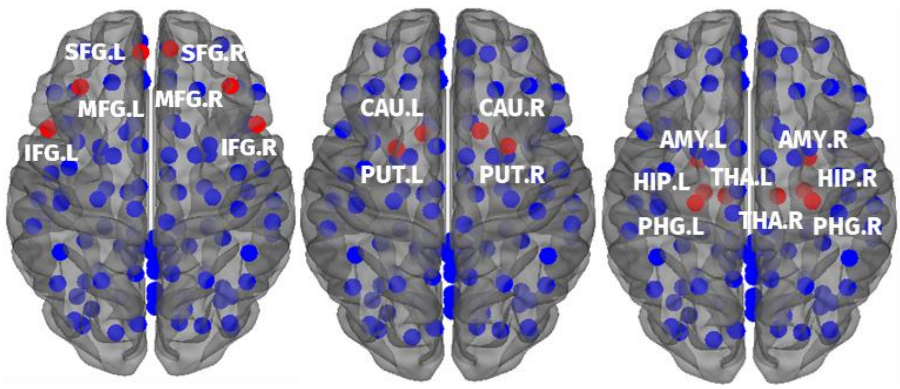


Figure 5.3: Location of brain regions relevant to an ADHD-related diagnosis, comprising the frontal lobe regions (left image), the basal ganglia (central image) and limbic structures (right image).

Basal ganglia and limbic structures

The basal ganglia are a set of subcortical regions that are in charge of motor and non-motor functions, several cognitive functions and emotional processing, with the striatum serving as the entry point to the basal ganglia, which is composed by the caudate nucleus (CAU), the putamen (PUT) and nucleus accumbens [133,134]. Throughout the years, researchers have found reduced volume or altered shapes in basal ganglia, as well as in limbic areas like amygdala (AMY), thalamus (THA) and hippocampus (HIP), in ADHD individuals when compared to healthy controls [135,136,137]. Dysfunction of the basal ganglia and fronto-striatal circuit, which includes the connections between the basal ganglia striatum, thalamus, and prefrontal regions, along with the limbic structures, are traditionally implied as one of the key interactions in ADHD pathophysiology. This circuit is linked to motivation and reward processing, where individuals with ADHD are able to concentrate during interesting activities, but are challenged by routine and everyday tasks. In addition to the structural alterations of these regions involved in ADHD, FC studies have also shown that the connectivity in these brain regions and circuits is affected. Recent works from [134] and [138] have reported that changes in the FC of striatum regions, in particular the caudate nucleus and the putamen, are related to ADHD groups. The study from [139] found an increased FC activity when analyzing the prefrontal cortex and the striatum, structures of the fronto-striatal circuit, among ADHD individuals. A hypoactivation of the FC activity in amygdala and hippocampus was also revealed, with these two regions showing abnormal FC activity in [140] study. Another work, from [141], examined the FC of the thalamus and two regions of the striatum, the caudate and putamen, showing an increase of FC in both structures of ADHD individuals. A meta-analysis conducted by [142] study, reinforced that the fronto-striatal regions, along with limbic regions, present indeed an abnormal FC in individuals with ADHD, supporting the previous studies statements.

All these regions involved in motivation and reward processing dysfunctions, were considered as relevant in the LRP analysis performed using the ConnectomeCNN model, corroborating the findings

presented above from diverse studies. The prefrontal regions were already identified in the previous LRP analysis, when the influence of this area on ADHD was presented, so the main focus will be on the basal ganglia and limbic system structures. Starting with basal ganglia, which comprise the striatum regions, the LRP analysis did not considered the nucleus accumbens as a relevant region for a diagnosis related to ADHD, since this region is not present in the AAL atlas used in this work. The caudate and putamen were considered, with the first being seen as relevant in the FC matrix computed with the correlation BCorrU method, and the latter considered as one of the most relevant brain regions by the FC matrix of the BCohF1 method. In relation to the limbic regions, the amygdala was identified by the LRP analysis as one of the most relevant brain regions for a diagnosis ADHD-related, namely in the two transfer entropy based methods, BTEU and BTED. The thalamus, another region belonging to the limbic system, was also seen as one of the most relevant regions, with its relevance being present when using FC matrices computed with the traditionally used correlation, which in this case are the BCorrU and BCorrD methods. The hippocampus joins the previous regions as relevant for distinguish ADHD individuals from typically developed controls, being considered as relevant by the LRP analysis when using BCorrU method.

Interestingly, the LRP also considered the parahippocampal gyrus (PHG), a region of the brain that surrounds the hippocampus and is part of the limbic system, as relevant, in the FC matrix of the BH2D method. FC abnormalities in this region have been associated to individuals with ADHD, demonstrated by few studies [130], as well as structural changes [137]. The location in the human brain of the basal ganglia and limbic structures considered the most relevant for a diagnosis related to ADHD, from the LRP analysis with the different FC matrices, are shown in the images on the center and on the right in figure 5.3, respectively.

Default-mode and cognitive control networks

A special network involved in ADHD related diagnosis, along with prefrontal brain regions, is the default-mode network (DMN). This network is a large-scale brain network that is well known for being active when the brain is at rest, becoming a hot topic of research with the objective of find and characterize dysfunctions in order to discover biomarkers for these brain connectivity abnormalities, since it is a brain network involved in many neurological and neuropsychiatric disorders [143]. DMN comprises the posterior cingulate gyrus (PCG), the precuneus (PCUN), medial prefrontal cortex, the medial, lateral, and inferior parietal cortex. One deficit present in ADHD is the response inhibition, where the subject can't prevent spontaneous and inappropriate responses, being this directly related to the activation of DMN instead of its suppression, contributing to a decreased task performance. Besides, interactions between DMN and cognitive control network, which includes the anterior cingulate cortex, the supplementary motor area (SMA), the posterior parietal cortex, the dorsolateral prefrontal cortex, and the inferior frontal junction, are important in ADHD pathophysiology. These interactions are activated when processes like working memory and inhibitory control happen [128,144]. Moreover, DMN and the cognitive control network functions are antagonists, when the levels of attention arise, the cognitive control network increases and diminishes the activation of the DMN. This suppression of DMN has been shown to be weaker in individuals with ADHD, indicating a disruption in the normal relationship between DMN and cognitive control network, which may be related to the neural mechanisms that lead to impairment of working memory [144]. In recent years, another region of brain has been revealed to be involved in the response inhibition, the postcentral gyrus (PoCG), located in somatosensory cortex, being this region associated to an abnormal connectivity with the precuneus [145]. Previous works have demonstrated increased FC in postcentral gyrus [145], [146], [147], as well as differences in the structure of the postcentral gyrus [137], finding that individuals with ADHD present

greater involvement in the processing of motor sensory information, which results in an impaired response inhibition by these individuals.

In this LRP analysis with ConnectomeCNN model, presented in table 5.7, many DMN brain regions were seen as relevant for an ADHD related diagnosis by the different statistical metrics used. Included in these brain regions are the precuneus in BCohF1 and BMITU methods, the posterior cingulate gyrus in BCorrD, BCohW1, BCohW2, BMITU, BTEU and BTED methods, the medial prefrontal cortex, in this AAL atlas is represented by the anterior cingulate gyrus (ACG), was found in BH2U method. Together with DMN regions, several cognitive control network were considered as relevant by the LRP technique in the ConnectomeCNN model. These regions comprise the supplementary motor area in BH2U method, regions of the posterior parietal cortex like the inferior parietal lobule (IPL), which are considered by BCohF1, BH2U and BTEU methods. Brain regions from the inferior frontal junction, involving the inferior frontal gyrus and the precentral gyrus (PreCG), were also identified by the LRP analysis, with the first region being seen as relevant by BCorrD, BCohF1, BCohF2 and BH2D methods, and the second region by BCorrD method as well, along with the BMITD1 and BMITD2 methods. Regarding the dorsolateral prefrontal cortex, which lies in middle frontal gyrus, the statistical methods responsible for considering this region as relevant, are the same discussed above in the frontal lobe brain regions. This LRP analysis was also able to reveal postcentral gyrus as one of the most relevant brain regions related to an ADHD diagnosis, being captured by two mutual information based methods, the BMITD1 and BMITD2. The location in the human brain of the DMN and cognitive control network regions considered the most relevant for a diagnosis related to ADHD, from the LRP analysis with the different FC matrices, are shown on the left and right images in figure 5.4, respectively.

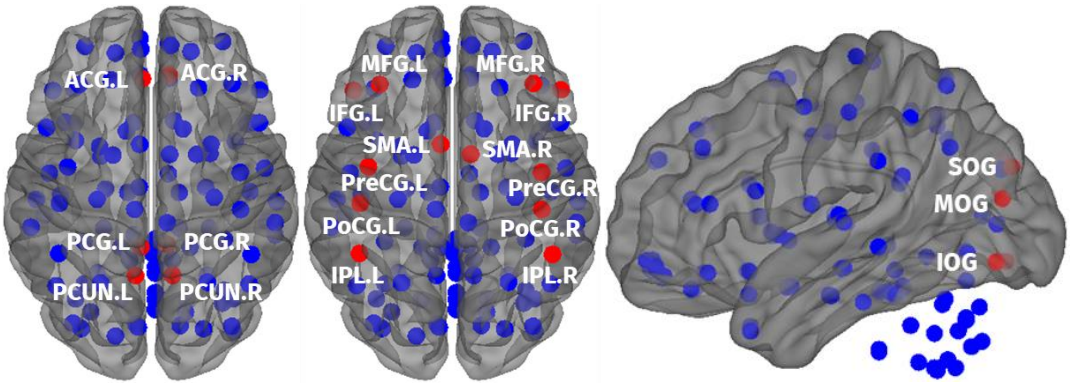


Figure 5.4: Location of brain regions relevant to an ADHD-related diagnosis, comprising the default-mode network (left image), the cognitive control network, including the postcentral gyrus (central image), and occipital cortex regions (right image).

The brain regions considered by the LRP analysis of this study are in agreement with the results obtained by other studies, including the study from [128], where the posterior cingulate gyrus, medial superior frontal gyrus and right inferior parietal lobule from ADHD individuals showed lower connectivity values in comparison with the same regions in controls. In another study [148], it was found that ADHD subjects showed decreased FC in anterior cingulate cortex, posterior cingulate cortex, lateral prefrontal cortex and precuneus. In addition to these studies, a more recent work performed in [149] examined the resting-state brain networks that differ most between normal subjects and subjects with ADHD, showing a decreased FC activity in DMN and in cognitive control network, indicating that the DMN is strongly related to the pathological basis of impaired response inhibition in ADHD.

Occipital cortex

The occipital cortex, in addition to processing visual information, is involved in the movement perception and in cognitive functions, being also a key region for working memory control, along with the DMN [128]. This brain region was also identified as important in the mechanism of working memory dysfunction in ADHD in [128] study, where the FC was increased in these regions when ADHD children were compared with normal controls, affecting the development of working memory by making this process slower. This study is in accordance with what was found in [129] and [150], where the occipital regions had an altered FC, showing an ADHD related functional abnormalities in these regions. The use of LRP with the ConnectomeCNN model in ADHD-200 dataset corroborates with the findings presented by previous studies. Revealed that the inferior (IOG), middle (MOG) and superior occipital gyrus (SOG) are involved in an ADHD-related diagnosis, being captured by the BCohW1, BCohW2, BH2U, BMITU, BMITD1 and BTEU statistical methods, as demonstrated in table 5.7. The image on the right in figure 5.4 shows the location in the human brain of the regions of the occipital cortex considered to be the most relevant for a diagnosis related to ADHD, from the LRP analysis with the different FC matrices.

Ventral and dorsal attention networks

ADHD is also characterized by symptoms of inattention, where two major brain systems are responsible for these processes, being those the ventral and dorsal attention networks. Ventral attention network (VAN) is associated with the orientation of attention when triggered by unexpected stimuli. VAN comprises the opercular part of the inferior frontal gyrus (IFGoper), the anterior cingulate gyrus, and the temporo-parietal junction, comprising the separation between the superior temporal gyrus and the middle temporal gyrus, and the inferior parietal lobule, this being divided into supramarginal (SMG) and angular gyrus (ANG). On the other hand, dorsal attention network (DAN) is prominently involved in voluntary and sustained control of attention, comprising the intraparietal sulcus, which separates the parietal lobe into superior (SPL) and inferior, the frontal eye field, which is located at the intersection of the middle frontal gyrus and the precentral gyrus [151,152]. The influence of the DAN in ADHD is quite evident in several studies, such as [153], [154] and [155] works, where it was found that the DAN among healthy individuals and individuals with ADHD presented a decrease in their FC activity, something confirmed more recently in [156] and [157] studies. Regarding the VAN, studies have found inconsistent results in terms of whether there is an increase or decrease in activity in this network. For example, the studies from [153] and [154], reported a decreased FC in VAN in ADHD adults and in ADHD children, respectively, however the latter found no alterations in the VAN of adults with ADHD. These discoveries were also present in [149] most recent work, confirming a reduced FC activity in children with ADHD, relative to healthy controls. On the other hand, the work from [158] found FC hyperactivation in both adults and children's ADHD groups, with adults showing greater FC activation, with the same increased FC in the of VAN ADHD children and adolescents reported in the study [159]. More recently in work [145], no significant changes were seen in the VAN of ADHD individuals. VAN's network suppression is fundamental to avoid the attentional shifting towards stimuli not related to the current task, with its hyperactivity being related to the distractibility symptoms of ADHD, although the contribution of this network in ADHD is not yet fully understood, while dysfunctions in DAN are clearly present [145]. Despite these findings, neither of the two networks exert a more important role in attention processes individually, but rather work together in a dynamic control of attention towards a specific goal [145,151].

The regions from the VAN and DAN involved in attention tasks mentioned above, are distinguished by the LRP analysis with ConnectomeCNN model as relevant for an ADHD-related diagnosis. From the regions present in VAN, even though the functions of this network in ADHD are not fully explicit, some

of these were considered as relevant for ADHD diagnosis, starting with the regions present in temporo-parietal junction. The supramarginal gyrus, part of the inferior parietal lobule, was considered relevant for an ADHD-related diagnosis in several statistical methods, being those the BCorrD, BCohW1, BCohF2, BCohW2 and all mutual information based methods. The other part of the inferior parietal lobule, the angular gyrus, was present in the LRP analysis of BCohW2, BH2U, and BMITD1 statistical methods. Regarding the remaining regions of the temporo-parietal junction, namely the temporal regions, were present in many statistical methods, such as the two correlation based methods, BCohF1, BCohF2, BCohW2, BH2U2, BMITU, BMITD2, and the two transfer entropy based methods. The opercular part of the inferior frontal gyrus is the portion of frontal lobe that overlaps the insula, and it was analyzed as relevant in BCorrD, BCohF1, BCohF2 and BH2D methods, while the anterior cingulate gyrus was only seen as a relevant region involved in ADHD in statistical method BH2U. Regions that are part of the DAN were also identified as relevant in the LRP analysis. Although there was no superior parietal lobule findings in LRP analysis, this technique was capable of identify the inferior parietal lobule as one of the regions more relevant for a diagnosis related to ADHD, mainly in BCohF1, BH2Ua and BTEU methods. Brain regions from the frontal eye field, which encompass around the middle frontal gyrus and the precentral gyrus, are present in LRP relevant regions. The middle frontal gyrus was seen as one of the most relevant brain regions for an ADHD-related diagnosis in BCohF1, BCohW2, BCohF2, BH2D, and all the mutual information based methods, with the precentral gyrus being captured by the BCorrD, BMITD1 and BMITD2 methods. The location in the human brain of the VAN and DAN regions considered the most relevant for a diagnosis related to ADHD, from the LRP analysis with the different FC matrices, are shown on the left and center images in figure 5.5, respectively.

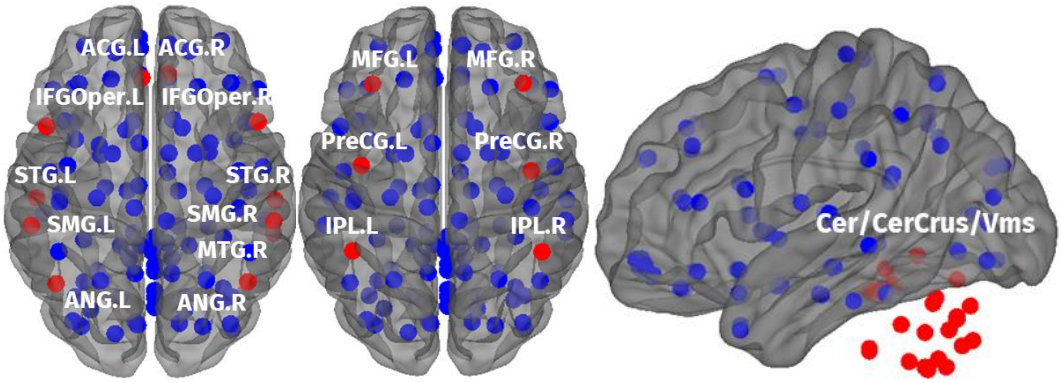


Figure 5.5: Location of brain regions relevant to an ADHD-related diagnosis, comprising the ventral attention network (left image), the dorsal attention network (central image) and cerebellum with its constituent vermis (right image).

Cerebellum

Recently, studies have also suggested the cerebellum as an active participant of this dorsal attention network [160]. The study performed in [161] found a decrease in the volume of the posterior cerebellar vermis and in the volume of the posterior inferior lobe in posterior vermis of individuals with ADHD, with a normal anterior portion and posterior superior lobe, respectively, demonstrating that the cerebellum is somehow involved in the voluntary attention system. Along with its relationship towards attention impairments, cerebellum has been widely associated with poor motor abilities symptoms in ADHD, in addition to impairments of regions involving motor control skills such as the supplementary motor area and the putamen, discussed previously. These poor motor abilities can affect 30-50% of children with ADHD, mostly in posture, walking, and balance capabilities [162]. Structural studies were the first to reveal consistent findings regarding alterations of cerebellar structures, more specifically reduced volume of the total cerebellum, cerebellar lobules and several portions of vermis, emphasizing

the role of the cerebellum and its constituents in ADHD [132,162]. The same findings of reductions in the volumes of cerebellum and its subregions, vermis and cerebellar lobules, were reported in the works of [163] and [164], with this reduction being correlated with ADHD, more specifically with the attentional and motor problems that arise from this diseased. In addition to the structural changes, an abnormal FC in the cerebellum and its regions was also verified. In [128] study, the FC intensity between both sides of the cerebellum was found to be increased in individuals with ADHD compared to what was observed in normal controls, reporting an indirect effect of this increase in the structure of the cerebellum, which is probable to intensify the ADHD symptoms and cause an attention impairment. Another study outcome confirmed the involvement of cerebellum as one of the brain regions with the most discriminative FC connections for in ADHD diagnosis, confirming its importance in ADHD abnormalities [138].

The LRP analysis performed with the ConnectomeCNN model considered the cerebellum (Cer/CerCrus), as well as the vermis (Vms), an integral structure of the cerebellum, two of the brain regions with most presence and some of the highest relevance values in FC matrices from different statistical metrics, in ADHD-related diagnosis, as seen in table 5.7. From these results, it is possible to observe that in the majority of these statistical methods used, with the exception of the BH2U, BMITD1 and BMITD2 methods, cerebellum and its vermis were the regions considered as the most relevant in a diagnosis related to ADHD. The results from this LRP analysis are in line with what has been revealed by the previous studies presented. The image on the right in figure 5.5 shows the location in the human brain of the regions of the cerebellum and vermis considered to be the most relevant for a diagnosis related to ADHD, from the LRP analysis with the different FC matrices.

Visual and auditory attention processing

Much of human attention comes from the processing of visual and auditory information, which, when not properly processed, can directly influence attention tasks from the main attention systems mentioned previously. Several regions located in the occipital lobe, such as the cuneus (CUN), lingual gyrus (LING) and calcarine cortex (CAL), as well as the fusiform gyrus (FFG), are responsible for the direct processing of visual information and form the visual attention system These regions maintain the attention levels and prevent distractions from undesirable stimuli, with the inability to inhibit unwanted stimuli as one of the main symptoms of ADHD [146]. Studies have been reporting FC abnormalities between ADHD and healthy individuals in these brain regions, namely in [130], [140], [145], [146], and more recently in [138] work. In addition to the FC abnormalities, findings on structural alterations are found in [137] study. All of these regions mentioned were considered relevant for a diagnosis related to ADHD in the LRP analysis performed with the ConnectomeCNN model. The cuneus and the lingual gyrus were present in the BCorrU method, the latter being also present in the BMITD2 method, with the fusiform gyrus being captured by the BCohW1 and BTED methods. The calcarine cortex was the region of the brain, among those involved in these mechanisms of visual attention, the region most present in statistical methods, being seen as relevant by the BCorrU, BCohF2, and both the transfer entropy methods, BTEU and BTED.

Not only dysfunctions of the visual attention system can be crucial in the process of impaired attention, but so is the auditory attention system, affecting the maintenance of individuals' attention and leading to their disturbance by external stimuli to the intended task. The auditory attention system is located in primary auditory cortex, which comprises portions of the superior temporal gyrus and the transverse temporal gyri, better known as Heschl's gyrus (HES). Although it is not yet fully understood, several studies have been reporting findings regarding the influence of auditory cortex in ADHD individuals, where deficits in this area can lead to significant high-order impairments, as presented in

[165] and [166] works. The work in [158] found enhanced FC in the auditory network when comparing ADHD and healthy controls groups, with other studies revealing in specific regions of this network, namely the increased FC activity in the superior temporal gyrus from a study performed by [145]. It was also found an increased FC activity in Heschl's gyrus in [166] and [167] works, with this increase in activity being related to a greater ability to be distracted by external stimuli to the task. The use of LRP analysis with the ConnectomeCNN model, regarding the regions considered relevant for a diagnosis of ADHD, was able to corroborate the findings of these studies presented. The superior temporal gyrus was already presented in the results of the LRP analysis before, while the Heschl's gyrus was considered as one of the most relevant brain regions only by the BTEU method. The location in the human brain of the visual and auditory attention processing regions considered the most relevant for a diagnosis related to ADHD, from the LRP analysis with the different FC matrices, are shown on the left image in figure 5.6.

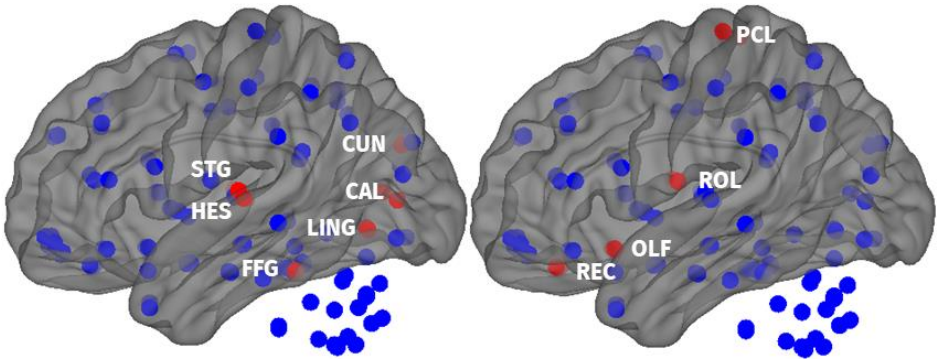


Figure 5.6: Location of brain regions involved in visual and auditory attention processing (left image) and other brain regions, such as the rectus gyrus, olfactory gyrus, Rolandic operculum and paracentral lobule, (right image) considered relevant for an ADHD-related diagnosis.

Rectus gyrus and olfactory gyrus

Studies are reporting several other brain regions beyond the most prominent ones, a majority of those discussed above. A brain region that is also involved in the diagnosis related to ADHD when LRP analysis is performed is the rectus gyrus (REC). The anterior surface of the orbital part of the frontal lobe is composed by the rectus gyrus, with this region being affected in ADHD, as verified by the studies from [168] and [169], observing an altered gray matter in individuals with ADHD when compared with healthy controls. Moreover, recent FC studies like [129], [138] and [140], have found rectus gyrus FC as one of the most discriminant brain regions in ADHD-related diagnosis. Together with rectus gyrus is the olfactory gyrus (OLF), also part of the anterior surface of the orbital part of the frontal lobe, since abnormalities this brain region have been contributing to ADHD, namely in working memory impairments, as demonstrated by [170] and [171] works. Along with the previous studies, FC abnormalities were also identified in olfactory gyrus in [129] and [138], revealing this region as one of the brain regions with the most discriminative power between ADHD and control individuals. The results achieved by the LRP analysis with the ConnectomeCNN model corroborate with the findings presented in the studies above, where the rectus gyrus was considered as a relevant brain region in ADHD-related diagnosis by several statistical methods, including the BCohF2, the h^2 and mutual information based methods. Regarding the olfactory gyrus, only one statistical method was involved in the ability of the LRP technique to consider this region as one of the ten most relevant for a diagnosis related to ADHD, this being the BH2D method.

Rolandic operculum and paracentral lobule

Another region that is beginning to be associated to ADHD problems is the Rolandic operculum (ROL), also known as the subcentral gyrus, which unifies the precentral and postcentral gyrus, and it is responsible for the control of emotions, language and speech, as well as in part of motor execution [172]. Together with the Rolandic operculum, the paracentral lobule (PCL) has also been linked to ADHD in recent findings, serving as a connector between the precentral and postcentral gyrus, controlling motor and sensory functions [173]. Some more modern studies have reported FC abnormalities in the Rolandic operculum, when comparing this region in individuals with ADHD and healthy controls, such as [125], [129] and [138]. Works by [138] and [174] showed alterations in the FC activity of paracentral lobule, and by [175] and [176] observed structural alterations in this region. These findings can be observed by the LRP analysis performed in this work and shown in table 5.7, where the Rolandic operculum was considered relevant for a diagnosis related to ADHD by the BMITD2 method. The paracentral lobule was considered relevant for ADHD diagnosis by more statistical methods, these being the BCohW2, BMITU, BMITD1, BMITD2 and the BTEU methods. As they are connected to precentral and postcentral gyrus, two regions with evidence of their influence on ADHD, as demonstrated in this LRP analysis with ConnectomeCNN model and previous studies, one might think that these may somehow affect the Rolandic operculum and the paracentral lobule, requiring further studies to assess the extent of this effect. The image on the right in figure 5.6 shows the location in the human brain of the rectus gyrus, olfactory gyrus, Rolandic operculum and paracentral lobule, considered to be the most relevant for a diagnosis related to ADHD, from the LRP analysis with the different FC matrices.

5.3.2 - LRP analysis with ConnectomeCNN-Autoencoder model

The same LRP analysis procedure is conducted when the classification for each individual FC matrix computed, from the set of statistical metrics used, is performed with the ConnectomeCNN-Autoencoder model. Figure 5.7 demonstrates a comparison between the mean FC matrix among all subjects present in the test set, computed through the BCorrU method, in the image on the right, and the respective visual heatmap result of the LRP analysis, in the image on the left. The LRP analysis is done for the FC matrices computed using the remaining statistical metrics, which are shown in figure A.3 of the Appendix.

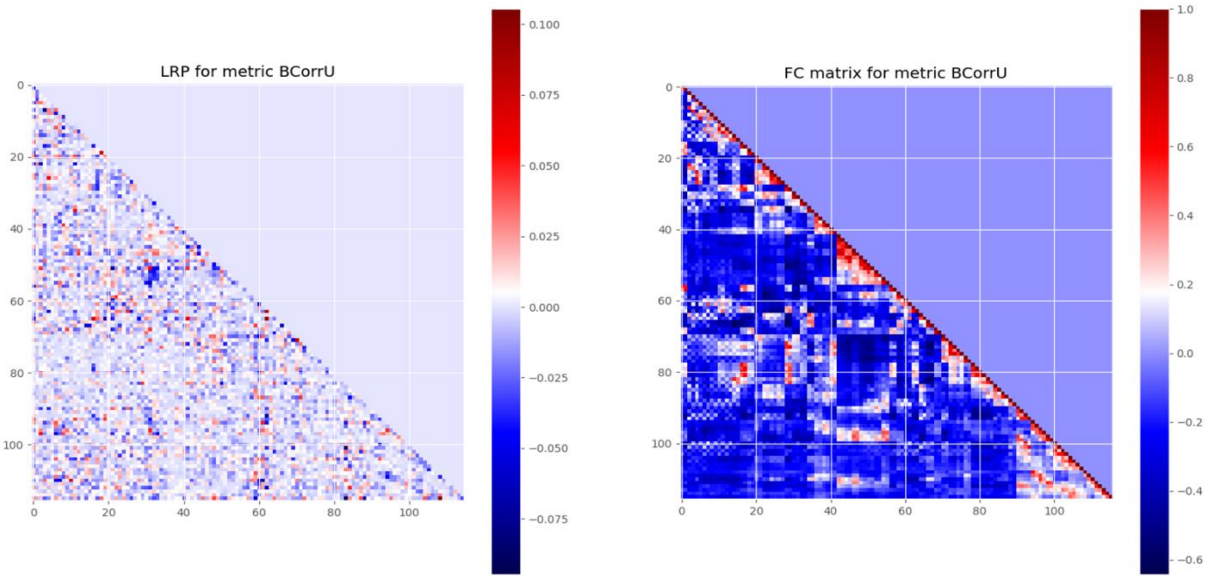


Figure 5.7: Heatmap of the Layer-wise Relevance Propagation analysis for the Functional Connectivity matrix computed with the undirected bivariate correlation method (left image) of the respective original Functional Connectivity matrix computed with the same method (right image) with the ConnectomeCNN-Autoencoder model, using the mean between the Functional Connectivity matrices of all subjects for this statistical method.

From the LRP analysis performed with the ConnectomeCNN-Autoencoder model, table 5.8 is presented, providing the ten most relevant brain regions related to an ADHD diagnosis, from the most relevant to the smallest among these ten regions, for each statistical metric FC matrix. The respective relevance values for these brain regions are shown in table 5.8 as well, exactly with the same order as in table 5.8.

Frontal lobe

By examining the LRP results for the ConnectomeCNN-Autoencoder model, it is possible conclude that, similarly to what was found earlier with the ConnectomeCNN model, the prefrontal cortex regions are the brain regions with the highest frequency in terms of being considered relevant for a diagnosis related to ADHD, among the FC matrices computed through the different statistical metrics, supporting the findings from FC and structural studies presented. The LRP analysis considered the superior frontal gyrus as relevant in several FC matrices of different statistical metrics, among which are the BCorrD, BCohF1, BCohF2, BCohW2, BH2D, and all mutual information based methods. Regarding the middle frontal gyrus as one of the most relevant regions for an ADHD-related diagnosis, it was seen by the LRP in BCorrD, BCohW1, BCohF2, BCohW2, BMITD2, and both transfer entropy methods, BTEU and BTED. The inferior frontal gyrus, in spite of being slightly less present as a relevant brain region compared to the other two regions of the prefrontal cortex, was considered relevant for a diagnosis related to ADHD in various statistical metrics of FC matrices, such as the BCohF1, BCohF2, BCohW2, BMITU, BMITD2 and BTED methods.

Basal ganglia and limbic structures

When the basal ganglia and the fronto-striatal circuit, together with the limbic structures, were analyzed by the LRP in conjunction with the ConnectomeCNN-Autoencoder model, they were found as relevant brain regions for an ADHD-related diagnosis in various statistical metrics of FC matrices. From the LRP analysis performed, regions of the striatum, which is composed by the caudate nucleus, the putamen and the nucleus accumbens, were included in table 5.8 among the most relevant regions for a diagnosis of ADHD. The caudate was considered as relevant in both correlation based methods, BCorrU and BCorrD, as well as in the BH2D method, with the putamen being seen as a relevant region for ADHD in BCohF1 and BTEU methods. The nucleus accumbens, a part of the striatum, is not considered in this analysis as this region is not part of the 116 AAL atlas.

Regarding the limbic system structures, involving the amygdala, thalamus and hippocampus, both were considered as relevant in this LRP analysis with the ConnectomeCNN-Autoencoder model. The amygdala was seen as relevant for an ADHD-related diagnosis in the FC matrix computed using BH2U method, and thalamus when used the FC matrices from BCorrD and BCohF1 methods. Concerning the hippocampus, this specific region was not considered as relevant for a diagnosis related to ADHD, but its surrounding region, the parahippocampal gyrus, which has been associated by some studies as a possible region affected by ADHD [130,137]. The results of the LRP analysis when used with the ConnectomeCNN-Autoencoder model corroborate the findings of several studies on the importance of these brain regions in the discrimination of individuals with ADHD.

Table 5.8a: Brain regions with greater impact on the others in an ADHD-related diagnosis, when using the ConnectomeCNN-Autoencoder model, and respective relevance values.

ConnectomeCNN-Autoencoder										
Method	**Brain Regions**									
BCorrU	Cer9.L	Cer6.L	Cer 4_5.R	Cer 4_5.L	TPOsup.R	STG.L	CAU.R	LING.L	PCG.L	OLF.L
Relevance Value	0.0015	0.0013	0.0012	0.00108	0.00100	0.00098	0.00089	0.00080	0.00068	0.0005
BCorrD	Vms 10	Vms 4_5	Cer 7b.L	CerCrus 2.L	THA.R	CAU.R	SOG.L	ORBmid.R	ORBmed.L	PreCG.L
Relevance Value	0.0012	0.0011	0.0011	0.00111	0.00097	0.00088	0.00083	0.00081	0.00080	0.0006
BCohF1	Vms 7	Cer 7b.L	THA.R	PUT.R	PUT.L	ANG.L	IPL.R	REC.L	SFGmed.R	IFGoper.L
Relevance Value	0.0010	0.0010	0.0010	0.00100	0.00087	0.00086	0.00075	0.00074	0.00072	0.0007
BCohW1	Cer4_5.R	CerCrus 1.L	HES.L	PoCG.R	FFG.R	FFG.L	IOG.L	IFGoper.L	ORBmid.L	MFG.L
Relevance Value	0.0021	0.0018	0.0017	0.00173	0.00166	0.00126	0.00126	0.00119	0.00112	0.0011
BCohF2	Cer 10.R	CerCrus 1.R	ITG.R	STG.R	STG.L	HES.L	REC.L	ORBmid.L	SFGmed.R	IFGtri.L
Relevance Value	0.0018	0.0016	0.0014	0.00131	0.00129	0.00126	0.00126	0.00115	0.00108	0.0010
BCohW2	Vms 10	Vms 9	IFG.R	FFG.R	FFG.L	MOG.L	PCG.L	SFGmed.R	MFG.L	ORBmed.R
Relevance Value	0.0016	0.0014	0.0013	0.00118	0.00116	0.00106	0.00104	0.00104	0.00102	0.0010
BH2U	Cer 10.L	Cer 3.R	ITG.R	MTG.L	IPL.L	SOG.R	AMY.R	ACG.L	REC.L	OLF.R
Relevance Value	0.0022	0.0018	0.0016	0.00131	0.00120	0.00115	0.00114	0.00109	0.00107	0.0010

Table 5.8b: Brain regions with greater impact on the others in an ADHD-related diagnosis, when using the ConnectomeCNN-Autoencoder model, and respective relevance values.

ConnectomeCNN-Autoencoder										
Method	**Brain Regions**									
BH2D	Cer 10.R	Cer 6.L	ITG.R	STG.L	CAU.R	SOG.L	SFGmed.R	SMA.R	SMA.L	ORBmed.L
Relevance Value	0.0016	0.0010	0.0009	0.00093	0.00088	0.00087	0.00070	0.00061	0.00060	0.0005
BMITU	Vms 7	Vms 3	CerCrus 2.L	MTG.R	STG.L	HES.R	PCUN.L	REC.L	ORBmed.L	IFGoper.R
Relevance Value	0.0022	0.0017	0.0011	0.00114	0.00113	0.00098	0.00094	0.00080	0.00070	0.0006
BMITD1	STG.R	STG.L	HES.R	IPL.L	CUN.R	CAL.L	MCG.L	SMA.L	ORBmed.L	Left PreCG
Relevance Value	0.0014	0.0014	0.0007	0.00063	0.00062	0.00052	0.00047	0.00044	0.00043	0.0003
BMITD2	Vms 4_5	STG.L	PCL.R	PHG.R	REC.L	ORBmed.R	SMA.R	IFGoper.R	ORBmid.R	ORBmed.L
Relevance Value	0.0009	0.0008	0.0007	0.00074	0.00074	0.00068	0.00068	0.00060	0.00056	0.0005
BTEU	Vms 9	Vms 1_2	HES.L	PUT.R	PCL.R	SPL.R	PCG.R	MCG.L	ORBmid.L	ROL.R
Relevance Value	0.0016	0.0014	0.0014	0.00136	0.00126	0.00123	0.00118	0.00115	0.00106	0.0010
BTED	Cer 9.R	STG.R	PCL.R	PCL.L	PCG.R	PCG.L	ORBmid.L	ROL.R	IFGorb.R	ORBmid.R
Relevance Value	0.0012	0.0011	0.0010	0.00090	0.00084	0.00082	0.00081	0.00067	0.00060	0.0006

Default-mode and cognitive control networks

DMN regions were identified as relevant for a diagnosis related to ADHD when used the LRP analysis with the ConnectomeCNN-Autoencoder model. Among these DMN regions, the precuneus is present in the FC matrix using BMITU method, the posterior cingulate gyrus using the BCorrU, BCohW2, and both transfer entropy methods, BTEU and BTED, with the medial prefrontal cortex, which is represented by the anterior cingulate gyrus in the 116 AAL atlas used, being considered as relevant for ADHD in BH2U method. As mentioned previously, cognitive control network regions are also proposed as implied in ADHD pathophysiology, and identified as relevant regions for ADHD diagnosis in the LRP analysis. The supplementary motor area was considered as relevant in the FC matrices of the BH2D, BMITD1 and BMITD2 methods, and a region of the posterior parietal cortex, like the inferior parietal lobule, was considered as relevant in BCohF1, BH2U, and BMITD1 methods.

Regions of the inferior frontal junction, also included in the cognitive control network, comprising the inferior frontal gyrus and the precentral gyrus, are seen as relevant regions for a diagnosis related to ADHD in FC matrices of the BCohF1, BCohF2, BCohW2, BMITU, BMITD2 and BTED methods, and in BCorrD and BMITD1 methods, respectively. The dorsolateral prefrontal cortex, located in the middle frontal gyrus, belongs to the cognitive control network, with the FC matrices of the statistical methods used considered as relevant for ADHD diagnosis being the same as the ones presented previously in frontal lobe regions. Regarding the postcentral gyrus, involved in the inhibition response deficits, it was seen as one of the most relevant brain regions related to a diagnosis of ADHD in the FC matrix of the BCohW1 method. These regions involved as relevant to a diagnosis of ADHD, through the LRP analysis with the ConnectomeCNN-Autoencoder model, are in line with what has been reported by several related studies, both in terms of FC and structural abnormalities.

Occipital cortex

The occipital cortex, despite including important regions in visual information processing, as will be discussed later, is also important in the control of working memory, together with DMN [128]. The regions of the occipital cortex, which play a key role in working memory, including the superior, middle and inferior occipital gyrus, were considered one of the most relevant regions of the LRP analysis with the ConnectomeCNN-Autoencoder model. These regions were found by the FC matrices of the BCorrD, BCohW1, and BCohW2 methods, as well as in both statistical methods from the h^2 metric, BH2U and BH2D.

Ventral and dorsal attention networks

One of the most important dysfunctions in ADHD pathophysiology is the inattention symptom, which is controlled by two major networks, the VAN and DAN. Of these two networks, VAN is the one with less consistent results regarding its impact on ADHD, even so presenting several regions in this LRP analysis with the ConnectomeCNN-Autoencoder model.

The temporo-parietal junction is one the most important area in the VAN, comprising the superior temporal gyrus and two regions of the inferior parietal lobule, the supramarginal and angular gyrus. The inferior parietal lobule and angular gyrus were identified by the LRP analysis as relevant regions regarding a diagnosis related to ADHD. Both regions were present in the FC matrix of the BCohF1 method, with the first one being also identified in the FC matrices of the BH2U and BMITD1 methods. Here in the LRP analysis obtained with the ConnectomeCNN-Autoencoder model from table 5.8, the remaining constituent of inferior parietal lobule, the supramarginal gyrus, was not considered relevant by any FC matrix. The temporal region of the temporo-parietal junction, the intersection between the

superior and middle temporal gyrus, was present as a relevant brain region in FC matrices of various statistical methods, such as the BCorrU, BCohF2, BH2U, BH2D, all mutual information based methods, BMITU, BMITD1, and BMITD2, as well as the BTED method. The VAN even encompasses more regions, such as the opercular part of the inferior frontal gyrus, considered as relevant for an ADHD-related diagnosis when using the BCohF1, BCohW1, BCohF2, BMITD1, BMITD2, and BTED methods, and the anterior cingulate cortex, only seen as important in the diagnosis of ADHD with the FC matrix of the BH2U method.

When it comes to the DAN, composed by the intraparietal sulcus, dividing the superior and inferior parietal lobules, and the frontal eye field, located between the middle frontal gyrus and the precentral gyrus, all these were identified by the LRP analysis with the model used, as presented in table 5.8. The intraparietal sulcus regions, superior and inferior parietal lobules, were identified as relevant regions by the LRP analysis when the FC matrices of the BTEU method, and the FC matrices of the BCohF1, BH2U and BMITD1 methods, respectively. The frontal eye field regions, including the middle frontal and precentral gyrus, were included among the brain regions with the most relevance in an ADHD-related diagnosis. The first was captured by the FC matrices of BCorrD, BCohW1, BCohF2, BCohW2, BMITD2, and both transfer entropy methods, BTEU and BTED, with the last to be identified in the FC matrices of the BCorrD and BMITD1 methods. These findings are in accordance with what has been reported in several studies, confirming the influence of DAN impairments on ADHD inattention and showing more evidence that VAN may also be involved in this problem.

Cerebellum

Looking at the table 5.8, the LRP analysis performed with the ConnectomeCNN-Autoencoder model considered the cerebellum and vermis as the brain regions with the highest relevance values in the majority of the FC matrices of the set of statistical methods used, except for the FC matrix of the BMITD1, for an ADHD-related diagnosis. As can be observed in the table mentioned, regions of the cerebellum were identified by LRP as the most relevant regions for a diagnosis of ADHD, among all regions considered in the analysis. This region was present in the FC matrix using the BCorrU method, having the four most relevant regions, in the FC matrix of the BCohW1, BCohF2, BH2U, and BH2D methods, with the two most relevant brain regions, as well as the most relevant brain region when using the FC matrix of the BTED method.

In addition to having the most important regions for a diagnosis related to ADHD, in the FC matrices of the BCorrD, BCohF1, and BMITU methods, cerebellum is present as one of the ten most important brain regions in this diagnosis. In the case of vermis, its presence as the most relevant brain region, among all the regions studied, was captured by the LRP in several FC matrices statistical methods, such as the FC matrices for the BCorrD, BCohW2, BMITU and BTEU methods, where two vermis structures were the two most relevant brain regions. In the FC matrices of the BCohF1 and BMITD2 methods, only one vermis region was considered as the most relevant region in ADHD diagnosis. The results achieved by the use of LRP analysis with the ConnectomeCNN-Autoencoder model showed a significant presence of the cerebellum and vermis for a diagnosis of ADHD, showing the importance of dysfunctions in these regions in the pathophysiology of the disease, reporting similar results to previous studies.

Visual and auditory attention processing

As discussed previously, the human attention is directly influenced by visual and auditory stimuli processing. The LRP analysis used along with the ConnectomeCNN-Autoencoder model displayed the importance of these visual processing brain regions in a diagnosis related to ADHD, as showed in table

5.8, being among the regions with the most relevance for that same diagnosis. The cuneus and calcarine cortex were seen as relevant brain regions when used the FC matrix of the BMITD1 method, with the lingual gyrus considered as relevant in the BCorrU method, while the fusiform gyrus was captured by the LRP analysis as relevant for ADHD diagnosis when used the FC matrices of the BCohW1 and BCohW2 methods.

Regions of the brain responsible for the auditory processing were also identified by the LRP analysis with the ConnectomeCNN-Autoencoder model as important regions for a diagnosis related to ADHD. Among these regions, the superior temporal gyrus was considered as a relevant region in FC matrices of the BCorrU, BCohF2, BH2D, and all mutual information based methods, BMITU, BMITD1, BMITD2, as well as the BTED method. The Heschl's gyrus was present in FC matrices of various statistical methods, such as the BCohW1, BCohF2, BMITU, BMITD1, and BTEU methods, showing its importance in the diagnosis of ADHD in this analysis. The results from this LRP analysis about the impact of visual and auditory processing brain regions are very positive, confirming several findings reported in previous works on these regions position in ADHD pathophysiology.

Rectus gyrus and olfactory gyrus

Recently, some brain regions that were not in the standard ADHD pathophysiology, as the ones discussed above, have started to emerge with evidence that their impairments were present in individuals with ADHD compared to typically developed controls. Among these regions is the rectus gyrus, identified by the LRP analysis with the ConnectomeCNN-Autoencoder model as a relevant region for a diagnosis related to ADHD, being considered in FC matrices of the BCohF1, BCohF2, BH2U, BMITD1, and BMITD2 methods. In addition, the olfactory gyrus is another brain region that has been associated with the pathophysiology of ADHD, which was seen as one of the most relevant brain regions in the LRP analysis with the ConnectomeCNN-Autoencoder model in FC matrices using the BCorrU and BH2U methods.

Rolandic operculum and paracentral lobule

Two other regions that have been shown to be affected by ADHD and were not in the standard ADHD pathophysiology are the Rolandic operculum and the paracentral lobule. In the LRP analysis with the ConnectomeCNN-Autoencoder model, paracentral lobule was highlighted as one of the most relevant regions for a diagnosis related to ADHD in FC matrices of the BMITD2, BTEU and BTED methods, with the Rolandic operculum being seen as relevant in both transfer entropy based methods, BTEU and BTED. All these findings from the LRP analysis seem to corroborate the results of recent studies presented on the influence of ADHD on dysfunctions in these regions.

5.3.3 – Overall View

From the results presented in tables 5.7 and 5.8, corresponding to the application of LRP analysis in conjunction with the ConnectomeCNN and ConnectomeCNN-Autoencoder models, respectively, it is possible to observe that the use of this XAI technique proved to be very similar in both models, in terms of the brain regions that were considered as relevant for an ADHD-related diagnosis. Since the results of the LRP analysis for both models used are practically identical, the figures of the location of brain regions considered relevant by this analysis will not be shown. A difference between the results obtained from the use of the two models concerns the superior parietal lobule, an element of DAN, a network involved in voluntary and sustained control of attention, in which attention impairments are a common symptom in subjects with ADHD. This region of the brain was considered as a relevant region in this

diagnosis when the LRP was used with the ConnectomeCNN-Autoencoder model, while in the ConnectomeCNN model only the inferior parietal lobule was found as relevant.

By using the FC matrices of the different statistical metrics in the LRP analysis, it was possible to assess how the use of statistical metrics in addition to the traditional linear metric used in FC studies, correlation coefficient, can actually provide more information, rather than the use of correlation alone. Looking at tables 5.7 and 5.8, it is noticeable in the LRP analysis using both models, that brain regions comprising the prefrontal cortex, an important area of dysfunction in the pathophysiology of ADHD, as well as the cerebellum and vermis, are present in the FC matrices of correlation-based methods and FC matrices of other methods. Despite this, most of the remaining brain regions that are shown by previous studies to be involved in ADHD are identified by LRP analysis using FC matrices of the different statistical metrics used.

From the above analysis on the regions of the brain that are more related to ADHD diagnosis, it demonstrates that the models used, ConnectomeCNN and ConnectomeCNN-Autoencoder models, together with the LRP technique, can reveal some of the brain circuits regions involved in ADHD pathophysiology. These findings are in accordance with structural and functional alterations reported in several previous studies presented.

6 – Conclusions

One of the main goals of the present dissertation was to study the use of different statistical metrics to compute the FC matrices using the time-series of the BOLD signals from two widely used datasets, the ABIDE-I and ADHD-200 datasets, which encompass subjects with ASD and ADHD, respectively, as well as healthy subjects. The FC data were used as input to two DNNs models, the modified ConnectomeCNN and the innovative ConnectomeCNN-Autoencoder models, each having their respective parameters fine-tuned in relation to the input data used, performing the classification task between healthy and diseased subjects for each FC matrix computed with the several statistical metrics studied. In addition, it was also intended to observe the effect of combining these FC matrices computed with the different statistical metrics, in a single FC multi-metric, on the classification performance of the models used. As the final objective of this dissertation, the XAI technique LRP was used to internally analyze the functioning of the DNNs models used in relation to their predictions, in order to overcome the black-box problem associated with these algorithms.

Regarding the performance of each model, the ConnectomeCNN and ConnectomeCNN-Autoencoder models showed quite similar classification performances when used in both datasets. There were, in fact, some differences in the FC matrices results for certain statistical metrics between the ConnectomeCNN and ConnectomeCNN-Autoencoder models when applied to the ADHD-200 dataset, where the results were slightly worse in the ConnectomeCNN-Autoencoder models, which may have been due the fact that this dataset classes are not evenly balanced, affecting the performance of the autoencoder. It was possible to conclude that, from the classification results of both datasets in both models, it is easier to distinguish between ASD and healthy subjects than between ADHD and healthy subjects.

The results from the individual FC matrices classifications revealed promising findings when it comes to using other statistical metrics to create the FC matrices, apart from the traditionally used correlation metric, as other statistical metrics like h^2 and mutual information, which are characterized for having the ability to consider non-linearities in signals, showed similar and in some cases better classification performances. This is more evident when used the FC matrix of the h^2 based methods, where the overall performance of the FC matrix of the best method based on h^2 metric is slightly better than the FC matrix of the best method based on correlation, while FC matrices based in mutual information methods, although it did not obtain a better classification than correlation, achieved performances very close to those of correlation-based methods. This emphasizes the importance of considering non-linearities present in the BOLD signals acquired, overcoming the natural limitation of correlation family metrics, validating its importance, as discussed in [84] and [85], and its complement to the use of correlation metrics. In this study, the results showed that FC matrices calculated using statistical metrics like coherence transfer entropy had the worst classification performance, requiring further studies to understand the capability of these metrics when applied to rs-fMRI data.

In the FC multi-metric approach, despite not improving the classification performance compared to the best individual FC matrix from the statistical metrics, achieving similar classification results, some interesting results were found, mainly in its ability extract the most important features from the different FC matrices combined. This can be verified by observing the measures used to evaluate the performance of the models, where in the majority of the experiments, using the ConnectomeCNN and ConnectomeCNN-Autoencoder models in both datasets, each model evaluation measure value for the FC multi-metric corresponds to the result of the respective measure of the best individual FC matrix,

supporting the evidence that this approach can select the best features of the FC matrices combined. The importance of these FC multi-metric findings is close to the studies of [7] and [97], where in these works an improvement in the classification was observed when a multi-metric was used, showing a positive impact of the use of this combination of statistical metrics. In the final part of this dissertation, the early use of the LRP technique from the iNNvestigate toolbox proved to be very successful in unraveling the black-box problem of DNNs, being applied to the FC matrices of ADHD-200 dataset subjects, it was able to reveal the regions of the brain that were most relevant to an ADHD-related diagnosis, with these LRP findings being supported by several previous studies.

For future work, some limitations of this dissertation should be addressed. In the present work, for both ABIDE-I and ADHD-200 datasets, the different subtypes of ASD and ADHD were not considered separately. Furthermore, for the group of subjects in these datasets, a demographic division was not made, in terms of age, gender, and also considering other relevant phenotypic information, such as IQ levels, handedness, among others. These subject's demographic and phenotypic information are extremely important in order to reduce the heterogeneity of these disorders, where each subtype is more related to a specific group age, genre, or other information. The consideration of this type of information may lead to improvements in the classification performance of the DL models used, which were below expectations, since it will allow them to better distinguish between the patterns of a disease subtype and the patterns of subjects considered healthy. This would affect not only the classification performance as well as the LRP analysis performed, making this analysis more reliable. Another limitation of this dissertation lies in the TR values associated with the acquisition of BOLD signals from rs-fMRI. Different TR's were used by the different image locations of the different subjects, being somehow interesting to try to understand if this difference in values can have any influence on the FC data, and later on the respective classification. Moreover, the improvement of the DL models used, as well as the investigation of new developments in other state-of-the-art DL models applied to FC studies, should not be discarded, as this area is in rapid evolution and with incredible progress.

Given the conclusions above mentioned, it is clear that the field of connectomics is far from being fully understood and mastered, with the FC studies being a valuable tool for analyzing neurological and neuropsychiatric disorders. It is also important to continue to explore and improve DL models in terms of the diagnosis ability of these disorders, consequently leading to the discovery of more reliable and/or new biomarkers, since these models have the capacity to deal with high-dimensionality data, which is the case of connectivity data, and allow to learn features from raw data without laborious handmade feature selection executed in tradition ML algorithms. The application of a FC multi-metric needs to continue to be more studied, as it demonstrates the ability to capture the most important features of each statistical metric combined, leading to a more complete analysis of the FC among the different brain regions BOLD signals due to different sources of information from different statistical metrics combined. The incorporation of XAI techniques, such as LRP and others, in DL algorithms, should be further investigated as it may be the missing piece, along with a reliable classification/diagnosis performance, for these valuable technologies to be implemented in clinical settings and assist healthcare professionals in their decisions, as well as provide them an understanding of how the model is behaving for a given prediction.

References

[1] Farahani, F. V., Karwowski, W., & Lighthall, N. R. (2019). Application of Graph Theory for Identifying Connectivity Patterns in Human Brain Networks: A Systematic Review. *Frontiers in Neuroscience*, 13, 585.

[2] Fornito, A., Zalesky, A., & Breakspear, M. (2015). The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3), 159–172.

[3] Van den Heuvel, M. P., & Hulshoff Pol, H. E. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, 20(8), 519–534.

[4] Fox, M. D., & Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, 8(9), 700–711.

[5] Smith, S. M., Vidaurre, D., Beckmann, C. F., Glasser, M. F., Jenkinson, M., Miller, K. L., Van Essen, D. C. (2013). Functional connectomics from resting-state fMRI. *Trends in Cognitive Sciences*, 17(12), 666–682.

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

[7] Meszlényi, R. J., Buza, K., & Vidnyánszky, Z. (2017). Resting State fMRI Functional Connectivity-Based Classification Using a Convolutional Neural Network Architecture. *Frontiers in Neuroinformatics*, 11, 61.

[8] Du, Y., Fu, Z., & Calhoun, V. D. (2018). Classification and Prediction of Brain Disorders Using Functional Connectivity: Promising but Challenging. *Frontiers in Neuroscience*, 12, 525.

[9] Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Muller, K.-R. (2021). Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*, 109(3), 247–278.

[10] Yousaf, T., Dervenoulas, G., & Politis, M. (2018). Advances in MRI Methodology. *International Review of Neurobiology*, 141, 31-76.

[11] Smitha, K., Akhil Raja, K., Arun, K., Rajesh, P., Thomas, B., Kapilamoorthy, T., & Kesavadas, C. (2017). Resting state fMRI: A review on methods in resting state connectivity analysis and resting state networks. *The Neuroradiology Journal*, 30(4), 305–317.

[12] Buxton, R. B. (2013). The physics of functional magnetic resonance imaging (fMRI). *Reports on Progress in Physics*, 76(9), 096601.

[13] Fox, M. D., & Greicius, M. (2010). Clinical Applications of Resting State Functional Connectivity. *Frontiers in Systems Neuroscience*, 4, 19.

[14] Vértes, P. E., & Bullmore, E. T. (2014). Annual Research Review: Growth connectomics - the organization and reorganization of brain networks during normal and abnormal development. *Journal of Child Psychology and Psychiatry*, 56(3), 299–320.

[15] Phinyomark, A., Ibanez-Marcelo, E., & Petri, G. (2017). Resting-State fMRI Functional Connectivity: Big Data Preprocessing Pipelines and Topological Data Analysis. *IEEE Transactions on Big Data*, 3(4), 415–428.

[16] Deco, G., & Kringelbach, ML. (2014). Great expectations: using whole-brain computational connectomics for understanding neuropsychiatric disorders. *Neuron*, 84(5), 892-905.

[17] Craddock, R. C., Tungaraza, R. L., & Milham, M. P. (2015). Connectomics and new approaches for analyzing human brain functional connectivity. *GigaScience*, 1(4), 1-12.

[18] Fornito, A., & Bullmore, E. T. (2015). Connectomics: A new paradigm for understanding brain disease. *European Neuropsychopharmacology*, 25(5), 733–748.

[19] Rossini, P. M., Di Iorio, R., Bentivoglio, M., Bertini, G., Ferreri, F., Gerloff, C., Hallett, M. (2019). Methods for analysis of brain connectivity: an IFCN-sponsored review. *Clinical Neurophysiology*, 130(10), 1833-1858.

[20] Wang, Z., Dai, Z., Gong, G., Zhou, C., & He, Y. (2014). Understanding Structural-Functional Relationships in the Human Brain. *The Neuroscientist*, 21(3), 290–305.

[21] Rykhlevskaia, E., Gratton, G., & Fabiani, M. (2008). Combining structural and functional neuroimaging data for studying brain connectivity: A review. *Psychophysiology*, 45(2), 173–187

[22] Damoiseaux, J. S. (2017). Effects of aging on functional and structural brain connectivity. *NeuroImage*, 160, 32–40.

[23] Agarwal, N., Tekes, A., Poretti, A., Meoded, A., & Huisman, T. (2017). Pitfalls in Diffusion-Weighted and Diffusion Tensor Imaging of the Pediatric Brain. *Neuropediatrics*, 48(05), 340–349.

[24] Salama, G. R., Heier, L. A., Patel, P., Ramakrishna, R., Magge, R., & Tsiouris, A. J. (2018). Diffusion Weighted/Tensor Imaging, Functional MRI and Perfusion Weighted Imaging in Glioblastoma-Foundations and Future. *Frontiers in Neurology*, 8, 660.

[25] Sun, Y., Yin, Q., Fang, R., Yan, X., Wang, Y., Bezerianos, A., … & Sun, J. (2014). Disrupted Functional Brain Connectivity and Its Association to Structural Connectivity in Amnestic Mild Cognitive Impairment and Alzheimer's Disease. *PLoS ONE*, 9(5), e96505.

[26] Lv, H., Wang, Z., Tong, E., Williams, L. M., Zaharchuk, G., Zeineh, M., Wintermark, M. (2018). Resting-State Functional MRI: Everything That Nonexperts Have Always Wanted to Know. *American Journal of Neuroradiology*, 39(8), 1390-1399.

[27] Tagliazucchi, E., & Chialvo, D. (2011). The collective brain is critical. *arXiv e-prints*, arXiv-1103.

[28] Bastos, A. M., & Schoffelen, J. M. (2016). A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls. *Frontiers in Systems Neuroscience*, 9, 175.

[29] Liégeois, R., Laumann, T. O., Snyder, A. Z., Zhou, J., & Yeo, B.T.T. (2017). Interpreting temporal fluctuations in resting-state functional connectivity MRI. *NeuroImage*, 163, 437–455.

[30] Mahadevan, A. S., Tooley, U. A., Bertolero, M. A., Mackey, A. P., & Bassett, D. S. (2021). Evaluating the sensitivity of functional connectivity measures to motion artifact in resting-state fMRI data. *Neuroimage*, 241, 118408.

[31] Asuero, A. G., Sayago, A., & González, A. G. (2006). The Correlation Coefficient: An Overview. *Critical Reviews in Analytical Chemistry*, 36(1), 41–59.

[32] Li, K., Guo, L., Nie, J., Li, G., & Liu, T. (2009). Review of methods for functional brain connectivity detection using fMRI. *Computerized Medical Imaging and Graphics*, 33(2), 131–139.

[33] Bowyer, S. M. (2016). Coherence a measure of the brain networks: past and present. *Neuropsychiatric Electrophysiology*, 2(1), 1-12.

[34] Lachaux, J.-P., Lutz, A., Rudrauf, D., Cosmelli, D., Le Van Quyen, M., Martinerie, J., & Varela, F. (2002). Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence. *Neurophysiologie Clinique/Clinical Neurophysiology*, 32(3), 157–174.

[35] Sankari, Z., Adeli, H., & Adeli, A. (2012). Wavelet Coherence Model for Diagnosis of Alzheimer Disease. *Clinical EEG and Neuroscience*, 43(4), 268–278.

[36] Lopes da Silva, F., Pijn, J. P., & Boeijinga, P. (1989). Interdependence of EEG signals: Linear vs. nonlinear Associations and the significance of time delays and phase shifts. *Brain Topography*, 2(1-2), 9–18.

[37] Pereda, E., Quiroga, R. Q., & Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals. *Progress in Neurobiology*, 77(1-2), 1–37.

[38] Wendling, F., Bartolomei, F., Bellanger, J. J., & Chauvel, P. (2001). Interpretation of interdependencies in epileptic signals using a macroscopic physiological model of the EEG. *Clinical Neurophysiology*, 112(7), 1201–1218.

[39] Na, S. H., Jin, S.-H., Kim, S. Y., & Ham, B.-J. (2002). EEG in schizophrenic patients: mutual information analysis. *Clinical Neurophysiology*, 113(12), 1954–1960.

[40] Steuer, R., Kurths, J., Daub, C. O., Weise, J., & Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18(Suppl 2), S231–S240.

[41] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6), 066138.

[42] Barnett, L., & Bossomaier, T. (2012). Transfer entropy as a log-likelihood ratio. *Physical Review Letters*, 109(13), 138105.

[43] Shovon, M. H. I., Nandagopal, N., Vijayalakshmi, R., Du, J. T., & Cocks, B. (2016). Directed Connectivity Analysis of Functional Brain Networks during Cognitive Activity Using Transfer Entropy. *Neural Processing Letters*, 45(3), 807–824.

[44] Vicente, R., & Wibral, M. (2014). Efficient Estimation of Information Transfer. *Directed Information Measures in Neuroscience*, 37-58. Springer, Berlin, Heidelberg.

[45] Vicente, R., Wibral, M., Lindner, M., & Pipa, G. (2010). Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of Computational Neuroscience*, 30(1), 45–67.

[46] Tehrani-Saleh, A., & Adami, C. (2020). Can Transfer Entropy Infer Information Flow in Neuronal Circuits for Cognitive Processing? *Entropy*, 22(4), 385.

[47] Yang, J., Gohel, S., & Vachha, B. (2020). Current methods and new directions in resting state fMRI. *Clinical Imaging*, 65, 47-53.

[48] De Reus, M. A., & van den Heuvel, M. P. (2013). The parcellation-based connectome: Limitations and extensions. *NeuroImage*, 80, 397–404.

[49] Eickhoff, S. B., Yeo, B. T., & Genon, S. (2018). Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, 19(11), 672-686.

[50] Arslan, S., Ktena, S. I., Makropoulos, A., Robinson, E. C., Rueckert, D., & Parisot, S. (2018). Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex. *NeuroImage*, 170, 5–30.

[51] Wang, Q., Chen, R., JaJa, J., Jin, Y., Hong, L. E., & Herskovits, E. H. (2015). Connectivity-Based Brain Parcellation. *Neuroinformatics*, 14(1), 83–97.

[52] Venkatesh, M., Jaja, J., & Pessoa, L. (2020). Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification. *NeuroImage*, 207, 116398.

[53] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

[54] Mahesh, B. (2020). Machine Learning Algorithms-A Review. *International Journal of Science and Research (IJSR)*.[Internet], 9, 381-386.

[55] Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., … Iyengar, S. S. (2018). A Survey on Deep Learning. *ACM Computing Surveys*, 51(5), 1–36.

[56] Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*, 2018, 1–13.

[57] Starship Knowledge. (2020). Perceptrons – These Artificial Neurons are the Fundamentals of Neural Networks. Available online at: https://starship-knowledge.com/neural-networks-perceptrons#What_does_the_learning_process_look_like [Accessed:15-08-2021].

[58] Aloysius, N., & Geetha, M. (2017). A review on deep convolutional neural networks. *2017 International Conference on Communication and Signal Processing (ICCSP)*, (pp. 0588-0592). IEEE.

[59] Shrestha, A., & Mahmood, A. (2019). Review of Deep Learning Algorithms and Architectures. *IEEE Access*, 7, 53040–53065.

[60] Park, Y. S., & Lek, S. (2016). Artificial neural networks: multilayer perceptron for ecological modeling. *Developments in environmental modelling* (Vol. 28, pp. 123-140). Elsevier.

[61] Seth, Yashu. (2018). A Disciplined Approach to Neural Network Hyper-Parameters – Paper Dissected. Available online at: https://yashuseth.blog/2018/11/26/hyper-parameter-tuning-best-practices-learning-rate-batch-size-momentum-weight-decay/.

[62] Medium. (2020). Getting Started With CNN- what are convolutional neural network?. Available online at: https://medium.com/@kinisanketh/getting-started-with-cnn-18c03efc7d06.

[63] AI Geek Programmer. (2019). Convolutional neural network 2: architecture. Available online at: https://aigeekprogrammer.com/convolutional-neural-network-image-recognition-part-2/.

[64] Lopez Pinaya, W. H., Vieira, S., Garcia-Dias, R., & Mechelli, A. (2020). Autoencoders. *Machine Learning*, 193–208.

[65] Bank, D., Koenigstein, N., & Giryes, R. (2020). Autoencoders. *arXiv preprint*, arXiv:2003.05991.

[66] M, H., & M.N, S. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2), 01–11.

[67] Dalianis, H. (2018). Evaluation Metrics and Evaluation. *Clinical Text Mining*, 45–53.

[68] Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint*, arXiv:1811.12808.

[69] Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Željko, & Milica, T. (2017). Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science*, 7(1), Pages: 39.

[70] Stąpor, K. (2017). Evaluating and comparing classifiers: Review, some recommendations and limitations. *International Conference on Computer Recognition Systems* (pp. 12-21). Springer, Cham.

[71] Japkowicz, N., & Shah, M. (2015). Performance Evaluation in Machine Learning. *Machine Learning in Radiation Oncology*, 41–56. Springer, Cham.

[72] Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., ... & Asadi, H. (2019). Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *American Journal of Roentgenology*, 212(1), 38-43.

[73] Berrar D. (2018) Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, Volume 1, Elsevier, pp. 542-545.

[74] Hagras, H. (2018). Toward Human-Understandable, Explainable AI. *Computer*, 51(9), 28–36.

[75] Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137-141.

[76] Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint*, arXiv:1708.08296.

[77] Buhrmester, V., Münch, D., & Arens, M. (2021). Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*, 3(4), 966-989.

[78] Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., & Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1), 16–18.

[79] London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, 49(1), 15–21.

[80] Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1–15.

[81] Xie, N., Ras, G., van Gerven, M., & Doran, D. (2020). Explainable deep learning: A field guide for the uninitiated. *arXiv preprint*, arXiv:2004.14545.

[82] Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Muller, K.-R. (Eds.). (2019). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. *Lecture Notes in Computer Science*.

[83] Kohlbrenner, M., Bauer, A., Nakajima, S., Binder, A., Samek, W., & Lapuschkin, S. (2020). Towards best practice in explaining neural network decisions with LRP. *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.

[84] Zhang, W., Muravina, V., Azencott, R., Chu, Z. D., & Paldino, M. J. (2018). Mutual Information Better Quantifies Brain Network Architecture in Children with Epilepsy. *Computational and Mathematical Methods in Medicine, 2018*.

[85] Aggarwal, R., & Ranganathan, P. (2016). Common pitfalls in statistical analysis: The use of correlation techniques. *Perspectives in Clinical Research*, 7(4), 187–190.

[86] Gao, Z. K., Liu, X. R., Ma, C., Ma, K., Gao, S., & Zhang, J. (2020). Functional alteration of brain network in schizophrenia: An fMRI study based on mutual information. *EPL (Europhysics Letters)*, 128(5), 50005.

[87] DSouza, A. M., Abidin, A. Z., Leistritz, L., & Wismüller, A. (2017). Exploring connectivity with large-scale Granger causality on resting-state functional MRI. *Journal of Neuroscience Methods*, 287, 68–79.

[88] Mäki-Marttunen, V., Diez, I., Cortes, J. M., Chialvo, D. R., & Villarreal, M. (2013). Disruption of transfer entropy and inter-hemispheric brain functional connectivity in patients with disorder of consciousness. *Frontiers in Neuroinformatics*, 7, 24.

[89] Diez, I., Erramuzpe, A., Escudero, I., Mateos, B., Cabrera, A., Marinazzo, D., ... & Diaz, J. M. C. (2015). Information Flow Between Resting-State Networks. *Brain Connectivity*, 5(9), 554.

[90] Kumar, S., Yoo, K., Rosenberg, M. D., Scheinost, D., Constable, R. T., Zhang, S., Li, C. R., & Chun, M. M. (2019). An information network flow approach for measuring functional connectivity and predicting behavior. *Brain and Behavior*, 9(8), e01346.

[91] Sun, F. T., Miller, L. M., & D'Esposito, M. (2004). Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data. *NeuroImage*, 21(2), 647–658.

[92] Thirion, B., Dodel, S., & Poline, J.-B. (2006). Detection of signal synchronizations in resting-state fMRI datasets. *NeuroImage*, 29(1), 321–327.

[93] Chang, C., & Glover, G. H. (2010). Time–frequency dynamics of resting-state brain connectivity measured with fMRI. *NeuroImage*, 50(1), 81–98.

[94] Damaraju, E., Allen, E., Belger, A., Ford, J., McEwen, S., Mathalon, D., Mueller, B., Pearlson, G., Potkin, S., Preda, A., Turner, J., Vaidya, J., Erp, T.V., & Calhoun, V. (2014). Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage: Clinical*, 5, 298-308.

[95] Yaesoubi, M., Miller, R. L., Bustillo, J., Lim, K. O., Vaidya, J., & Calhoun, V. D. (2017). A joint time-frequency analysis of resting-state functional connectivity reveals novel patterns of connectivity shared between or unique to schizophrenia patients and healthy controls. *NeuroImage: Clinical*, 15, 761–768.

[96] Al-Hiyali, M. I., Yahya, N., Faye, I., & Hussein, A. F. (2021). Identification of Autism Subtypes Based on Wavelet Coherence of BOLD FMRI Signals Using Convolutional Neural Network. *Sensors*, 21(16), 5256.

[97] Mohanty, R., Sethares, W. A., Nair, V. A., & Prabhakaran, V. (2020). Rethinking measures of functional connectivity via feature extraction. *Scientific Reports*, 10(1), 1-17.

[98] Valliani, A. A., Ranti, D., & Oermann, E. K. (2019). Deep Learning and Neurology: A Systematic Review. *Neurology and Therapy*, 8(2), 351–365.

[99] Li, H., Parikh, N. A., & He, L. (2018). A Novel Transfer Learning Approach to Enhance Deep Neural Network Classification of Brain Functional Connectomes. *Frontiers in Neuroscience*, 12, 491.

[100] Khosla, M., Jamison, K., Kuceyeski, A., & Sabuncu, M. R. (2018). 3D convolutional neural networks for classification of functional connectomes. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 137-145). Springer, Cham.

[101] Kuang, D., Guo, X., An, X., Zhao, Y., & He, L. (2014). Discrimination of ADHD Based on fMRI Data with Deep Belief Network. *Lecture Notes in Computer Science*, 225–232.

[102] Kim, J., Calhoun, V. D., Shim, E., & Lee, J. H. (2016). Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *NeuroImage*, 124(Pt A), 127-146.

[103] Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, 17, 16–23.

[104] Eslami, T., Mirjalili, V., Fong, A., Laird, A., & Saeed, F. (2019). ASD-DiagNet: A Hybrid Learning Approach for Detection of Autism Spectrum Disorder Using fMRI Data. *Frontiers in Neuroinformatics*, 13, 70.

[105] Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R., Zwicker, J., & Hamarneh, G. (2017). BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146, 1038-1049.

[106] Brown, C. J., Kawahara, J., & Hamarneh, G. (2018). Connectome priors in deep neural networks to predict autism. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (pp. 110-113). IEEE.

[107] Shahriman, W. N. S., Phang, C. R., Numan, F., & Ting, C. M. (2020). Classification of Brain Functional Connectivity using Convolutional Neural Networks. *IOP Conference Series: Materials Science and Engineering* (Vol. 884, No. 1, p. 012003). IOP Publishing.

[108] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint*, arXiv:1312.6034.

[109] Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint*, arXiv:1412.6806.

[110] Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision* (pp. 818-833). Springer, Cham.

[111] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE*, 10(7), e0130140.

[112] Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification. *Frontiers in Aging Neuroscience*, 11, 194.

[113] Yan, W., Plis, S., Calhoun, V. D., Liu, S., Jiang, R., Jiang, T. Z., & Sui, J. (2017). Discriminating schizophrenia from normal controls using resting state functional network connectivity: A deep neural network and layer-wise relevance propagation method. *2017 IEEE 27th international workshop on machine learning for signal processing (MLSP)* (pp. 1-6). IEEE.

[114] ABIDE Preprocessed Connectomes Project website. Visited on 30/03/2021. Available online at: http://preprocessed-connectomes-project.org/abide/.

[115] ABIDE Preprocessed Connectomes Project preprocessing website. Visited on 24/07/2021. Available online at: http://preprocessed-connectomes-project.org/abide/dparsf.html.

[116] ADHD-200 Preprocessed Connectomes Project website. Visited on 24/07/2021. Available online at: http://preprocessed-connectomes-project.org/adhd200/.

[117] Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D. S., & Craddock, R. C. (2017). The Neuro Bureau ADHD-200 Preprocessed repository. *NeuroImage*, 144, 275–286.

[118] Mijalkov, M., Kakaei, E., Pereira, J. B., Westman, E., Volpe, G., & Alzheimer's Disease Neuroimaging Initiative. (2017). BRAPH: a graph theory software for the analysis of brain connectivity. *PLoS ONE*, 12(8), e0178798.

[119] Wang, H. E., Bénar, C. G., Quilichini, P. P., Friston, K. J., Jirsa, V. K., & Bernard, C. (2014). A systematic framework for functional connectivity measures. *Frontiers in Neuroscience*, 8, 405.

[120] Autism Brain Imaging Data Exchange I- ABIDE I website. Visited on 24/07/2021. Available online at: http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html

[121] MULAN toolbox. 2016. Visited on 04/08/2021. Available online at: https://github.com/HuifangWang/MULAN.

[122] Towards Data Science. 2020. 12 Main Dropout Methods: Mathematical and Visual Explanation for DNNs, CNNs, and RNNs. Available online at: https://towardsdatascience.com/12-main-dropout-methods-mathematical-and-visual-explanation-58cdc2112293

[123] Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., ... & Kindermans, P. J. (2019). iNNvestigate Neural Networks!. *Journal of Machine Learning Research*, 20, 1-8.

[124] Mao, Z., Su, Y., Xu, G., Wang, X., Huang, Y., Yue, W., Sun, L., & Xiong, N.N. (2019). Spatio-temporal deep learning method for ADHD fMRI classification. *Information Sciences*, 499, 1-11.

[125] Aradhya, A. M. S., Joglekar, A., Suresh, S., & Pratama, M. (2019). Deep Transformation Method for Discriminant Analysis of Multi-Channel Resting State fMRI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 2556-2563.

[126] Chayer, C., & Freedman, M. (2001). Frontal lobe functions. *Current Neurology and Neuroscience Reports*, 1(6), 547-552.

[127] Hoffmann, M. (2013). The human frontal lobes and frontal network systems: an evolutionary, clinical, and treatment perspective. *International Scholarly Research Notices*, 2013, 892459.

[128] Jiang, K., Yi, Y., Li, L., Li, H., Shen, H., Zhao, F., ... & Zheng, A. (2019). Functional network connectivity changes in children with attention-deficit hyperactivity disorder: a resting-state fMRI study. *International Journal of Developmental Neuroscience*, 78, 1-6.

[129] Tang, Y., Wang, C., Chen, Y., Sun, N., Jiang, A., & Wang, Z. (2021). Identifying ADHD Individuals From Resting-State Functional Connectivity Using Subspace Clustering and Binary Hypothesis Testing. *Journal of Attention Disorders*, 25(5), 736-748.

[130] Wang, W., Hu, B., Yao, Z., Jackson, M., Liu, R., & Liang, C. (2013). Dysfunctional neural activity and connection patterns in attention deficit hyperactivity disorder: A resting state fMRI study. *The 2013 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-6). IEEE.

[131] Riaz, A., Asad, M., Alonso, E., & Slabaugh, G. (2020). DeepFMRI: End-to-end deep learning for functional connectivity and classification of ADHD using fMRI. *Journal of Neuroscience Methods*, 335, 108506.

[132] Krain, A. L., & Castellanos, F. X. (2006). Brain development and ADHD. *Clinical Psychology Review*, 26(4), 433–444.

[133] Durston, S., van Belle, J., & de Zeeuw, P. (2011). Differentiating frontostriatal and fronto-cerebellar circuits in attention-deficit/hyperactivity disorder. *Biological Psychiatry*, 69(12), 1178–1184.

[134] Sörös, P., Hoxhaj, E., Borel, P., Sadohara, C., Feige, B., Matthies, S., ... & Philipsen, A. (2019). Hyperactivity/restlessness is associated with increased functional connectivity in adults with ADHD: a dimensional analysis of resting state fMRI. *BMC Psychiatry*, 19(1), 1-11.

[135] Rubia K. (2018). Cognitive Neuroscience of Attention Deficit Hyperactivity Disorder (ADHD) and Its Clinical Translation. *Frontiers in Human Neuroscience*, 12, 100.

[136] Damiani, S., Tarchi, L., Scalabrini, A., Marini, S., Provenzani, U., Rocchetti, M., Oliva, F., & Politi, P. (2020). Beneath the surface: hyper-connectivity between caudate and salience regions in ADHD fMRI at rest. *European Child & Adolescent Psychiatry*, 1-13.

[137] Gehricke, J. G., Kruggel, F., Thampipop, T., Alejo, S. D., Tatos, E., Fallon, J., & Muftuler, L. T. (2017). The brain anatomy of attention-deficit/hyperactivity disorder in young adults - a magnetic resonance imaging study. *PLoS ONE*, 12(4), e0175433.

[138] Sun, Y., Zhao, L., Lan, Z., Jia, X. Z., & Xue, S. W. (2020). Differentiating Boys with ADHD from Those with Typical Development Based on Whole-Brain Functional Connections Using a Machine Learning Approach. *Neuropsychiatric Disease and Treatment*, 16, 691–702.

[139] Rosch, K. S., Mostofsky, S. H., & Nebel, M. B. (2018). ADHD-related sex differences in fronto-subcortical intrinsic functional connectivity and associations with delay discounting. *Journal of Neurodevelopmental Disorders*, 10(1), 1-14.

[140] Tang, Y., Li, X., Chen, Y., Zhong, Y., Jiang, A., & Wang, C. (2020). High-Accuracy Classification of Attention Deficit Hyperactivity Disorder With l2,1-Norm Linear Discriminant Analysis and Binary Hypothesis Testing. *IEEE Access*, 8, 56228-56237.

[141] Mills, K. L., Bathula, D., Dias, T. G., Iyer, S. P., Fenesy, M. C., Musser, E. D., Stevens, C. A., Thurlow, B. L., Carpenter, S. D., Nagel, B. J., Nigg, J. T., & Fair, D. A. (2012). Altered cortico-striatal-thalamic connectivity in relation to spatial working memory capacity in children with ADHD. *Frontiers in Psychiatry*, 3, 2.

[142] Norman, L. J., Carlisi, C., Lukito, S., Hart, H., Mataix-Cols, D., Radua, J., & Rubia, K. (2016). Structural and Functional Brain Abnormalities in Attention-Deficit/Hyperactivity Disorder and Obsessive-Compulsive Disorder: A Comparative Meta-analysis. *JAMA Psychiatry*, 73(8), 815–825.

[143] Fransson, P., & Marrelec, G. (2008). The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: Evidence from a partial correlation network analysis. *NeuroImage*, 42(3), 1178–1184.

[144] Mohan, A., Roberto, A. J., Mohan, A., Lorenzo, A., Jones, K., Carney, M. J., ... & Lapidus, K. A. (2016). Focus: the aging brain: the significance of the default mode network (DMN) in neurological and neuropsychiatric disorders: a review. *The Yale Journal of Biology and Medicine*, 89(1), 49.

[145] Lin, H., Lin, Q., Li, H., Wang, M., Chen, H., Liang, Y., Bu, X., Wang, W., Yi, Y., Zhao, Y., Zhang, X., Xie, Y., Du, S., Yang, C., & Huang, X. (2021). Functional Connectivity of Attention-Related Networks in Drug-Naïve Children With ADHD. *Journal of Attention Disorders*, 25(3), 377–388.

[146] Tang, C., Wei, Y., Zhao, J., & Nie, J. (2018). Different Developmental Pattern of Brain Activities in ADHD: A Study of Resting-State fMRI. *Developmental Neuroscience*, 40(3), 246–257.

[147] Lei, D., Du, M., Wu, M., Chen, T., Huang, X., Du, X., Bi, F., Kemp, G. J., & Gong, Q. (2015). Functional MRI reveals different response inhibition between adults and children with ADHD. *Neuropsychology*, 29(6), 874–881.

[148] Qiu, M. G., Ye, Z., Li, Q. Y., Liu, G. J., Xie, B., & Wang, J. (2011). Changes of brain structure and function in ADHD children. *Brain Topography*, 24(3-4), 243–252.

[149] Li, J., Joshi, A. A., & Leahy, R. M. (2020). A Network-Based Approach to Study of Adhd Using Tensor Decomposition of Resting State Fmri Data. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (pp. 1-5). IEEE.

[150] Wang, L., Zhu, C., He, Y., Zang, Y., Cao, Q., Zhang, H., Zhong, Q., & Wang, Y. (2009). Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder. *Human Brain Mapping*, 30(2), 638–649.

[151] Vossel, S., Geng, J. J., & Fink, G. R. (2014). Dorsal and Ventral Attention Systems: Distinct Neural Circuits but Collaborative Roles. *The Neuroscientist*, 20(2), 150-159.

[152] Salmi, J., Salmela, V., Salo, E., Mikkola, K., Leppämäki, S., Tani, P., ... & Alho, K. (2018). Out of focus–Brain attention control deficits in adult ADHD. *Brain Research*, 1692, 12-22.

[153] Cortese, S., Kelly, C., Chabernaud, C., Proal, E., Di Martino, A., Milham, M. P., & Castellanos, F. X. (2012). Toward systems neuroscience of ADHD: a meta-analysis of 55 fMRI studies. *The American Journal of Psychiatry*, 169(10), 1038–1055.

[154] Tomasi, D., & Volkow, N. D. (2012). Abnormal functional connectivity in children with attention-deficit/hyperactivity disorder. *Biological Psychiatry*, 71(5), 443–450.

[155] McCarthy, H., Skokauskas, N., Mulligan, A., Donohoe, G., Mullins, D., Kelly, J., Johnson, K., Fagan, A., Gill, M., Meaney, J., & Frodl, T. (2013). Attention network hypoconnectivity with default and affective network hyperconnectivity in adults diagnosed with attention-deficit/hyperactivity disorder in childhood. *JAMA Psychiatry*, 70(12), 1329–1337.

[156] Zhang, H., Zhao, Y., Cao, W., Cui, D., Jiao, Q., Lu, W., Li, H., & Qiu, J. (2020). Aberrant functional connectivity in resting state networks of ADHD patients revealed by independent component analysis. *BMC Neuroscience*, 21(1), 39.

[157] Guo, X., Yao, D., Cao, Q., Liu, L., Zhao, Q., Li, H., ... & Sun, L. (2020). Shared and distinct resting functional connectivity in children and adults with attention-deficit/hyperactivity disorder. *Translational Psychiatry*, 10(1), 1-12.

[158] Farrant, K., & Uddin, L. Q. (2015). Asymmetric development of dorsal and ventral attention networks in the human brain. *Developmental Cognitive Neuroscience*, 12, 165–174.

[159] Sanefuji, M., Craig, M., Parlatini, V., Mehta, M. A., Murphy, D. G., Catani, M., ... & de Schotten, M. T. (2017). Double-dissociation between the mechanism leading to impulsivity and inattention in attention deficit hyperactivity disorder: a resting-state functional connectivity study. *Cortex*, 86, 290-302.

[160] Brissenden, J. A., Tobyne, S. M., Osher, D. E., Levin, E. J., Halko, M. A., & Somers, D. C. (2018). Topographic Cortico-cerebellar Networks Revealed by Visual Attention and Working Memory. *Current Biology*, 28(21), 3364-3372.

[161] Durston, S., Hulshoff Pol, H. E., Schnack, H. G., Buitelaar, J. K., Steenhuis, M. P., Minderaa, R. B., Kahn, R. S., & van Engeland, H. (2004). Magnetic resonance imaging of boys with attention-deficit/hyperactivity disorder and their unaffected siblings. *Journal of the American Academy of Child and Adolescent Psychiatry*, 43(3), 332–340.

[162] Bruchhage, M., Bucci, M. P., & Becker, E. (2018). Cerebellar involvement in autism and ADHD. *Handbook of Clinical Neurology*, 155, 61–72.

[163] Stoodley C. J. (2014). Distinct regions of the cerebellum show gray matter decreases in autism, ADHD, and developmental dyslexia. *Frontiers in Systems Neuroscience*, 8, 92.

[164] Buckner R. L. (2013). The cerebellum and cognitive function: 25 years of insight from anatomy and neuroimaging. *Neuron*, 80(3), 807–815.

[165] Scheich, H., Brechmann, A., Brosch, M., Budinger, E., Ohl, F. W., Selezneva, E., Stark, H., Tischmeyer, W., & Wetzel, W. (2011). Behavioral semantics of learning and crossmodal processing in auditory cortex: the semantic processor concept. *Hearing Research*, 271(1-2), 3–15.

[166] Serrallach, B., Groß, C., Bernhofs, V., Engelmann, D., Benner, J., Gündert, N., Blatow, M., Wengenroth, M., Seitz, A., Brunner, M., Seither, S., Parncutt, R., Schneider, P., & Seither-Preisler, A.

(2016). Neural Biomarkers for Dyslexia, ADHD, and ADD in the Auditory Cortex of Children. *Frontiers in Neuroscience*, 10, 324.

[167] Rolls, E. T., Cheng, W., & Feng, J. (2021). Brain dynamics: the temporal variability of connectivity, and differences in schizophrenia and ADHD. *Translational Psychiatry*, 11(1), 1-11.

[168] Griffiths, K. R., Grieve, S. M., Kohn, M. R., Clarke, S., Williams, L. M., & Korgaonkar, M. S. (2016). Altered gray matter organization in children and adolescents with ADHD: a structural covariance Connectome study. *Translational Psychiatry*, 6(11), e947.

[169] Stevens, M. C., & Haney-Caron, E. (2012). Comparison of brain volume abnormalities between ADHD and conduct disorder in adolescence. *Journal of Psychiatry & Neuroscience : JPN*, 37(6), 389–398.

[170] Cockcroft K. (2011). Working memory functioning in children with attention-deficit/hyperactivity disorder (ADHD): A comparison between subtypes and normal controls. *Journal of Child and Adolescent Mental Health*, 23(2), 107–118.

[171] Crow, A., Janssen, J. M., Vickers, K. L., Parish-Morris, J., Moberg, P. J., & Roalf, D. R. (2020). Olfactory Dysfunction in Neurodevelopmental Disorders: A Meta-analytic Review of Autism Spectrum Disorders, Attention Deficit/Hyperactivity Disorder and Obsessive-Compulsive Disorder. *Journal of Autism and Developmental Disorders*, 50(8), 2685–2697.

[172] Triarhou L. C. (2021). Cytoarchitectonics of the Rolandic operculum: morphofunctional ponderings. *Brain Structure & Function*, 226(4), 941–950.

[173] Patra, A., Kaur, H., Chaudhary, P., Asghar, A., & Singal, A. (2021). Morphology and Morphometry of Human Paracentral Lobule: An Anatomical Study with its Application in Neurosurgery. *Asian Journal of Neurosurgery*, 16(2), 349–354.

[174] Mehren, A., Özyurt, J., Lam, A. P., Brandes, M., Müller, H., Thiel, C. M., & Philipsen, A. (2019). Acute Effects of Aerobic Exercise on Executive Function and Attention in Adult Patients With ADHD. *Frontiers in Psychiatry*, 10, 132.

[175] Saute, R., Dabbs, K., Jones, J. E., Jackson, D. C., Seidenberg, M., & Hermann, B. P. (2014). Brain morphology in children with epilepsy and ADHD. *PLoS ONE*, 9(4), e95269.

[176] Zou, H., & Yang, J. (2021). Exploring the Brain Lateralization in ADHD Based on Variability of Resting-State fMRI Signal. *Journal of Attention Disorders*, 25(2), 258–264.

Appendix A

Table A.1: Number of subjects in the ABIDE-I dataset from each imaging institution used in the study.

ABIDE I Institution	Number of subjects
California Institute of Technology	37
Kennedy Krieger Institute	39
University of Leuven	59
Ludwig Maximilian's University Munich	41
NYU Langone Medical Center	169
Oregon Health and Science University	23
Institute of Living at Hartford Hospital	24
University of Pittsburgh School of Medicine	36
Social Brain Lab, Groningen Institute of Neurosciences	15
San Diego State University	33
Stanford University	36
Trinity Centre of Health Sciences	44
University of California, Los Angeles	75
University of Michigan	113
University of Utah School of Medicine	61
Yale Child Study Center	48

Table A.2 Number of subjects in the ADHD-200 dataset from each imaging institution used in the study.

ADHD-200 Institution	Number of subjects
NeuroIMAGE	48
Kennedy Krieger Institute	83
Peking University	194
Washington University	59
NYU Langone Medical Center	216
Oregon Health and Science University	79
University of Pittsburgh School of Medicine	89

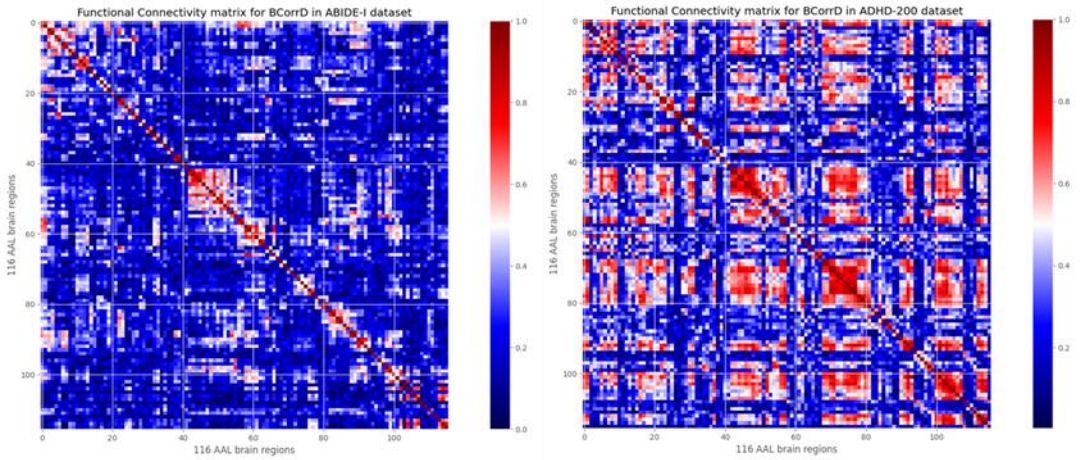
Table A.3: Repetition time for each imaging institution from ABIDE I dataset present in the study.

ABIDE I Institution	TR (seconds)
California Institute of Technology	2
Kennedy Krieger Institute	2.5
University of Leuven	0.0016
Ludwig Maximilian's University Munich	3
NYU Langone Medical Center	2
Oregon Health and Science University	2.5
Institute of Living at Hartford Hospital	1.5
University of Pittsburgh School of Medicine	1.5
Social Brain Lab, Groningen Institute of Neurosciences	2.2
San Diego State University	2
Stanford University	2

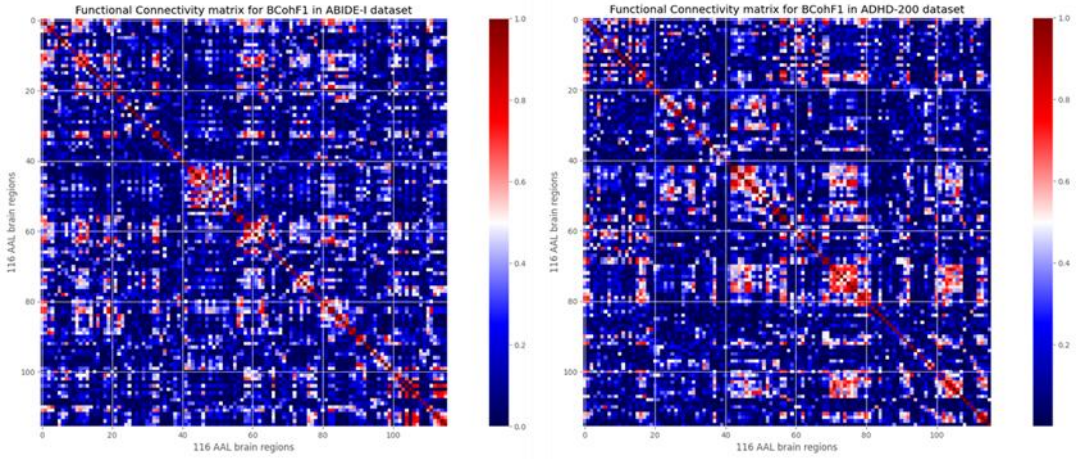
Trinity Centre of Health Sciences	2
University of California, Los Angeles	3
University of Michigan	2
University of Utah School of Medicine	2
Yale Child Study Center	2

Table A.4: Repetition time for each imaging institution from ADHD-200 dataset present in the study.

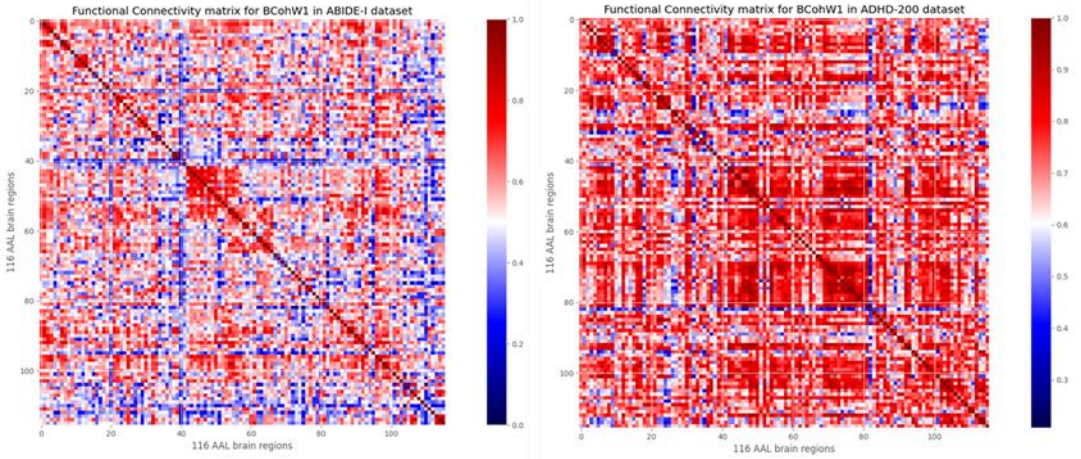
ADHD-200 Institution	**TR (seconds)**
NeuroIMAGE	1.960
Kennedy Krieger Institute	2.5
Peking University	2
Washington University	2.5
NYU Langone Medical Center	2
Oregon Health and Science University	2.5
University of Pittsburgh School of Medicine	1.5


Functional Connectivity matrix for BCorrD in ABIDE-I dataset
116 AAL brain regions
116 AAL brain regions

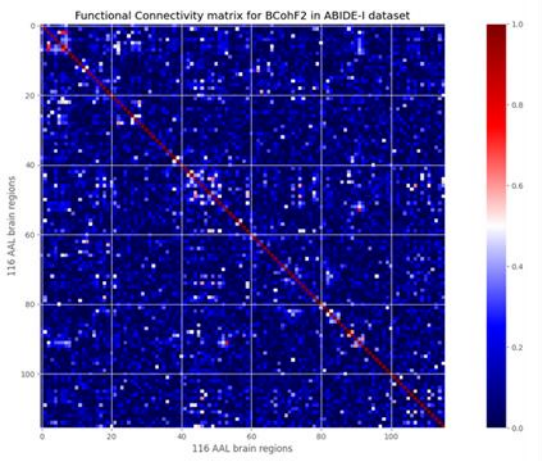
Functional Connectivity matrix for BCorrD in ADHD-200 dataset
116 AAL brain regions
116 AAL brain regions

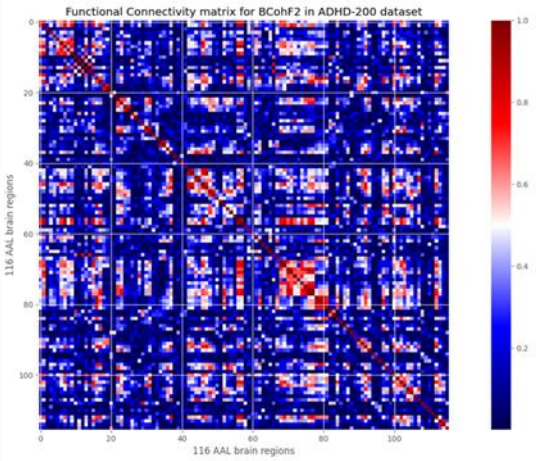

Functional Connectivity matrix for BCohF1 in ABIDE-I dataset
116 AAL brain regions
116 AAL brain regions

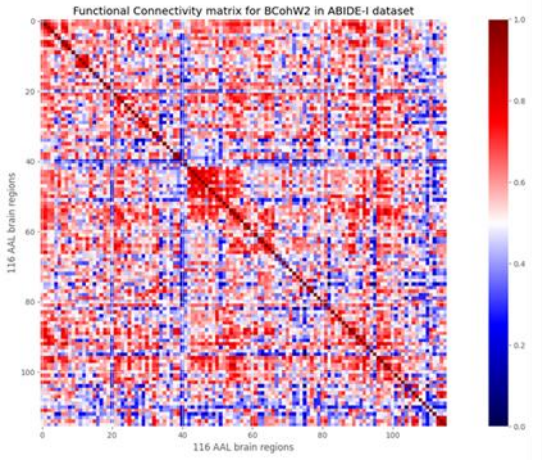
Functional Connectivity matrix for BCohF1 in ADHD-200 dataset
116 AAL brain regions
116 AAL brain regions

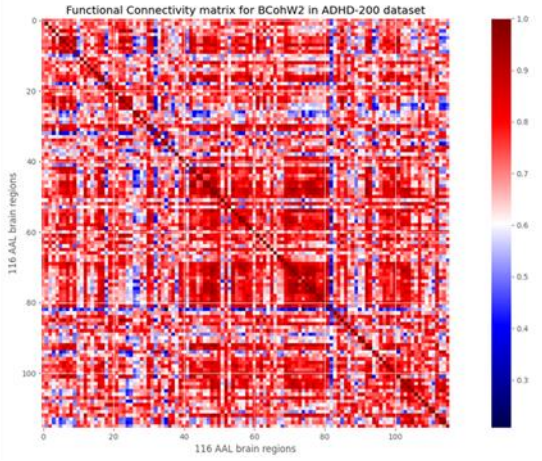

Functional Connectivity matrix for BCohW1 in ABIDE-I dataset
116 AAL brain regions
116 AAL brain regions

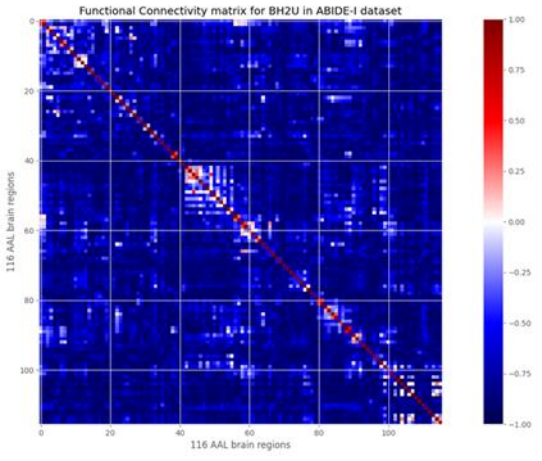
Functional Connectivity matrix for BCohW1 in ADHD-200 dataset
116 AAL brain regions
116 AAL brain regions

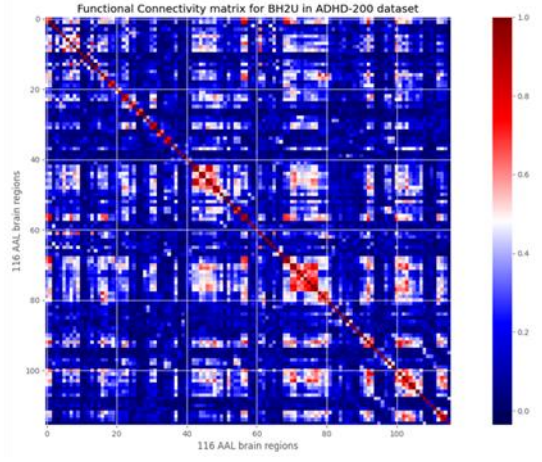

Functional Connectivity matrix for BCohF2 in ABIDE-I dataset
116 AAL brain regions
116 AAL brain regions

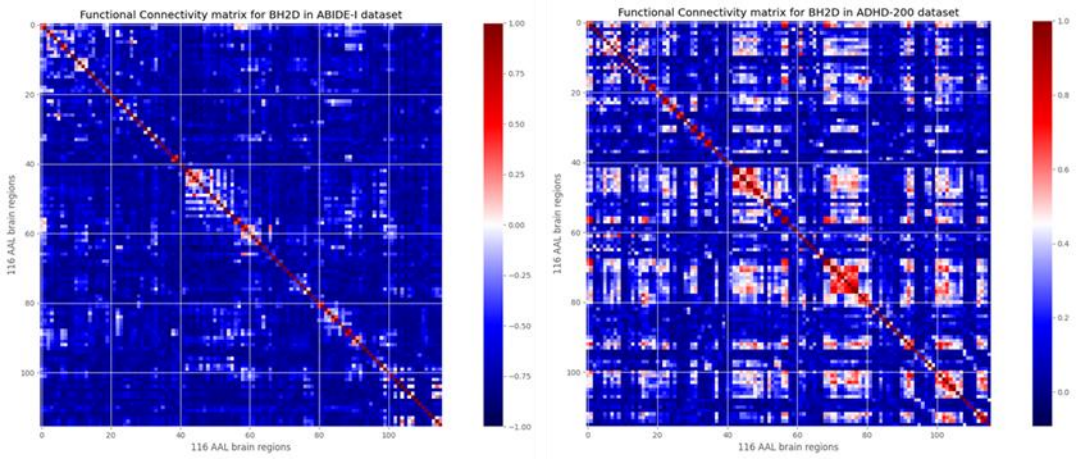

Functional Connectivity matrix for BCohF2 in ADHD-200 dataset
116 AAL brain regions
116 AAL brain regions

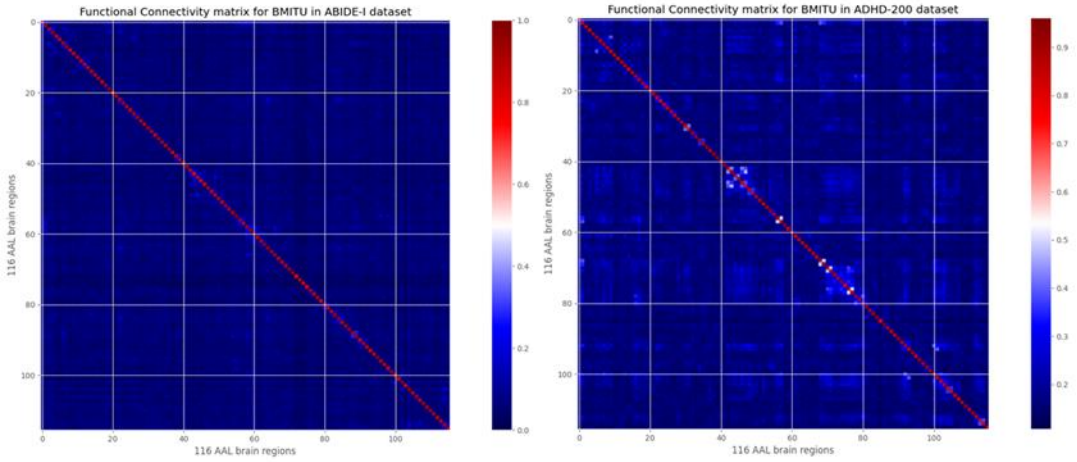

Functional Connectivity matrix for BCohW2 in ABIDE-I dataset
116 AAL brain regions
116 AAL brain regions

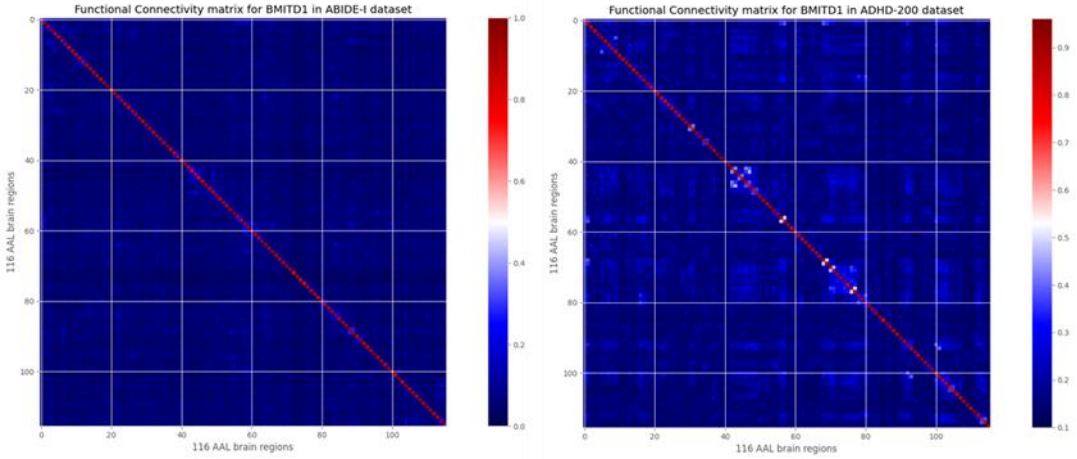

Functional Connectivity matrix for BCohW2 in ADHD-200 dataset
116 AAL brain regions
116 AAL brain regions


Functional Connectivity matrix for BH2U in ABIDE-I dataset
116 AAL brain regions
116 AAL brain regions


Functional Connectivity matrix for BH2U in ADHD-200 dataset
116 AAL brain regions
116 AAL brain regions


Functional Connectivity matrix for BH2D in ABIDE-I dataset
116 AAL brain regions
116 AAL brain regions
Functional Connectivity matrix for BH2D in ADHD-200 dataset
116 AAL brain regions
116 AAL brain regions


Functional Connectivity matrix for BMITU in ABIDE-I dataset
116 AAL brain regions
116 AAL brain regions
Functional Connectivity matrix for BMITU in ADHD-200 dataset
116 AAL brain regions
116 AAL brain regions


Functional Connectivity matrix for BMITD1 in ABIDE-I dataset
116 AAL brain regions
116 AAL brain regions
Functional Connectivity matrix for BMITD1 in ADHD-200 dataset
116 AAL brain regions
116 AAL brain regions

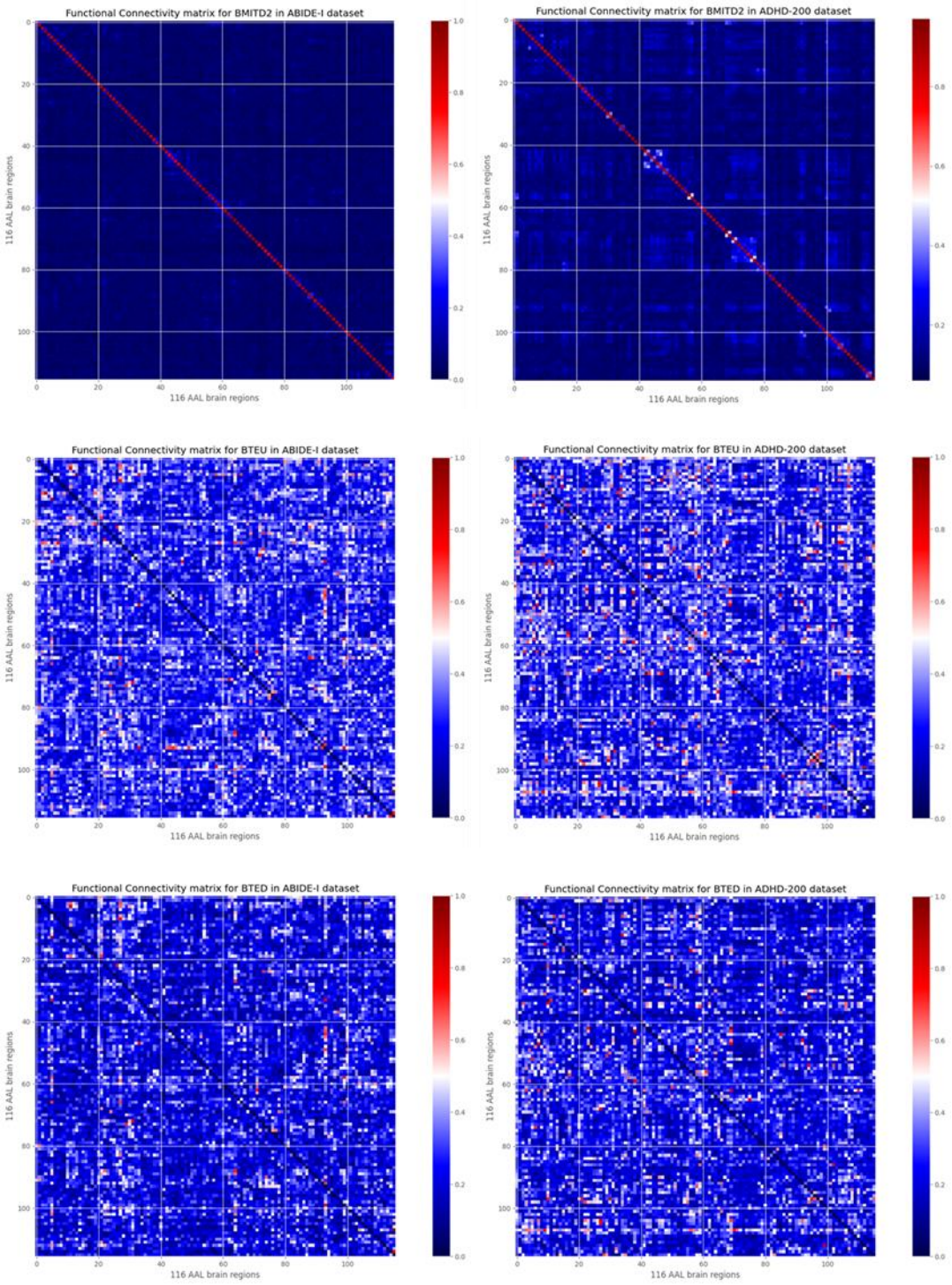
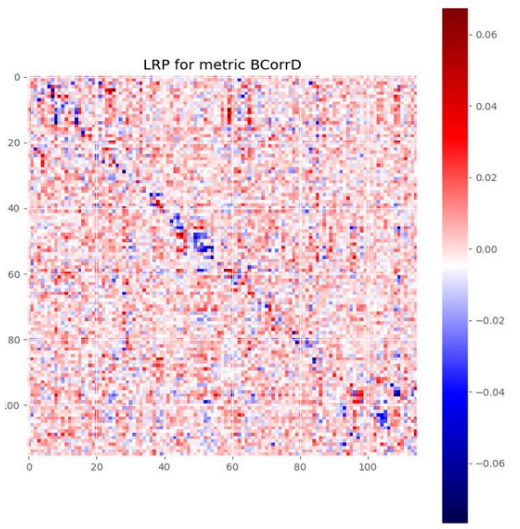
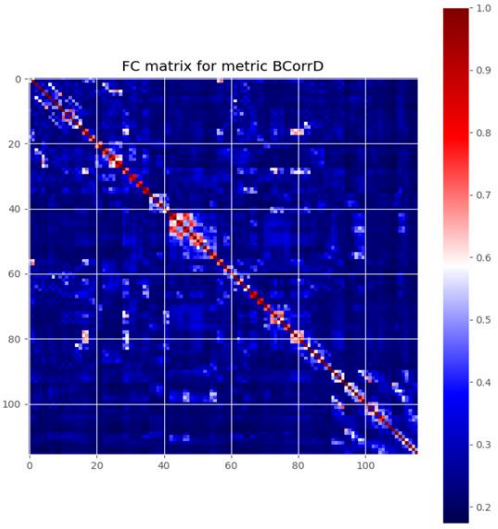
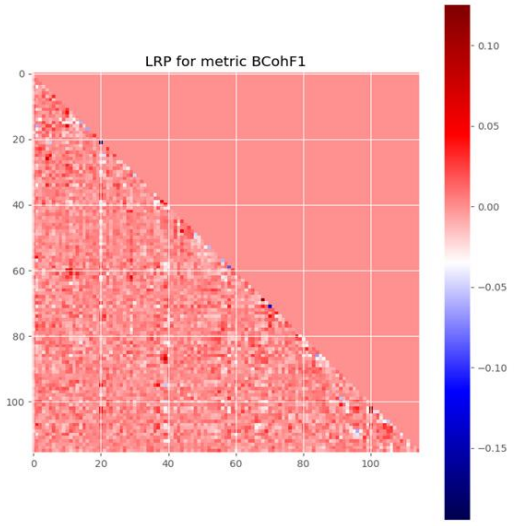
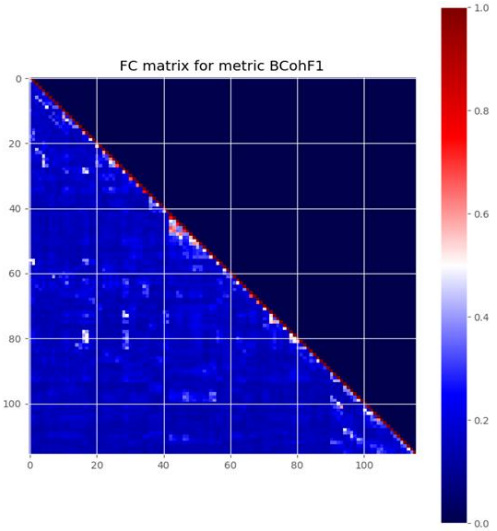


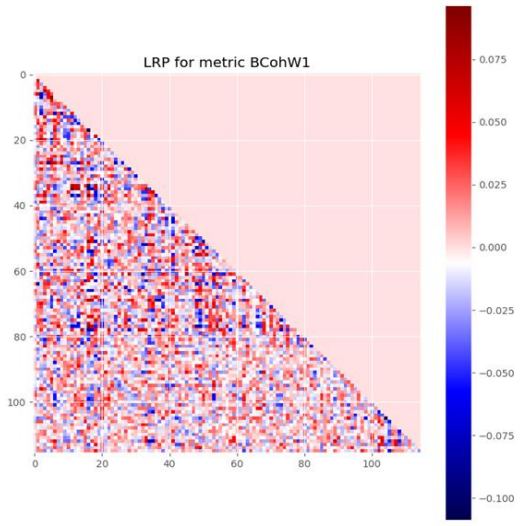
Figure A.1: Visualization of the Functional Connectivity matrices computed using the remaining statistical metrics methods chosen for this study, BCorrD, BCohF1, BCohW1, BCohF2, BCohW2, BH2U, BH2D, BMITU, BMITD1, BMITD2, BTEU and BTED (top to bottom), where the left image corresponds to a random subject from ABIDE-I dataset and the right image corresponds to a random subject from ADHD-200 dataset.

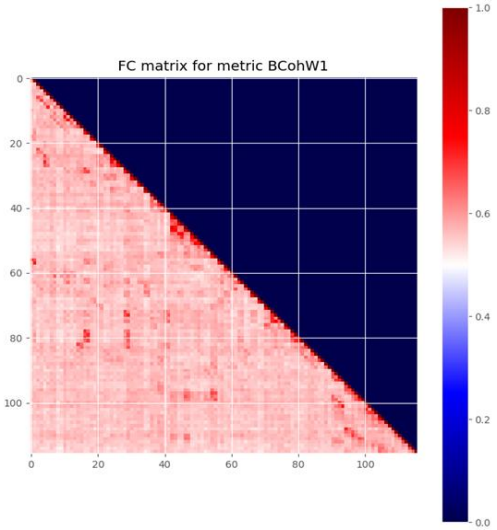

LRP for metric BCorrD

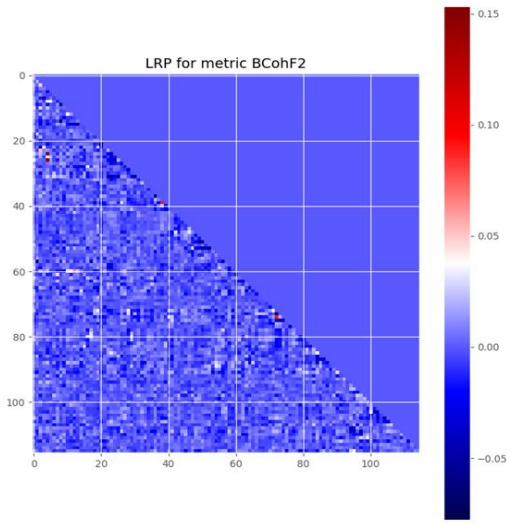

FC matrix for metric BCorrD

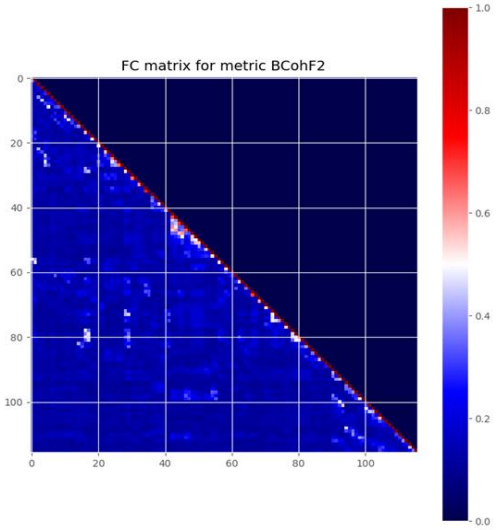

LRP for metric BCohF1

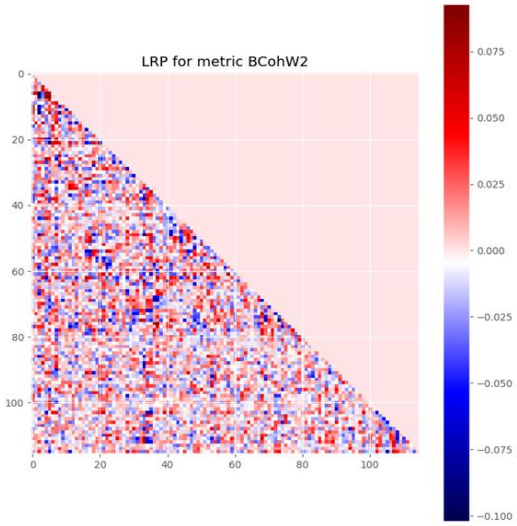

FC matrix for metric BCohF1

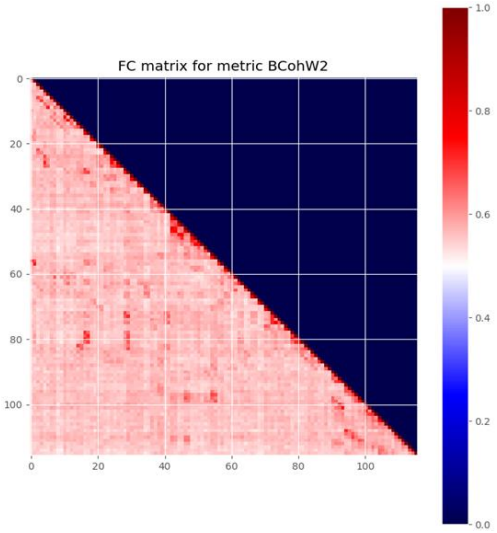

LRP for metric BCohW1

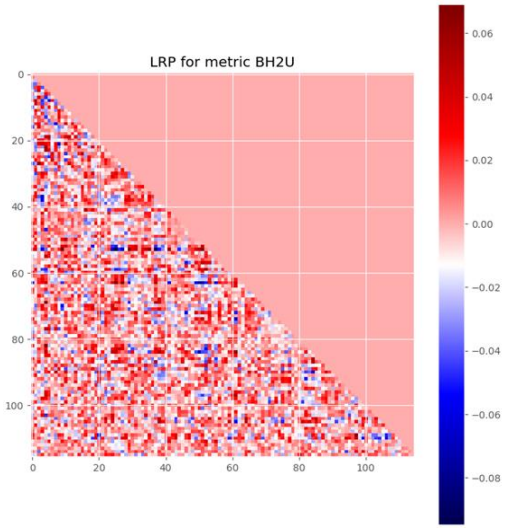

FC matrix for metric BCohW1

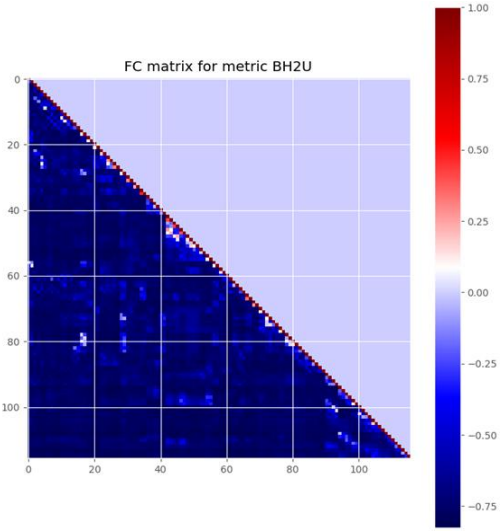

LRP for metric BCohF2

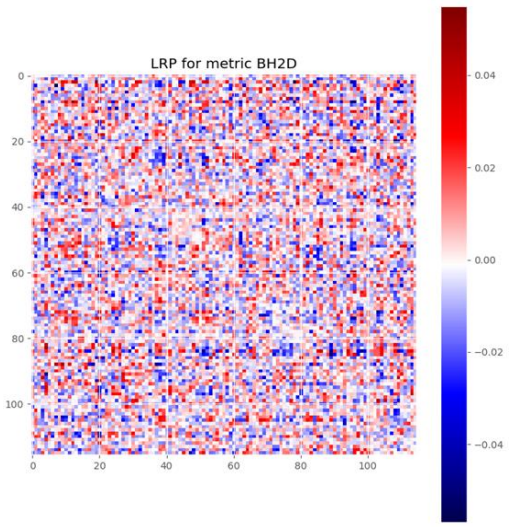

FC matrix for metric BCohF2

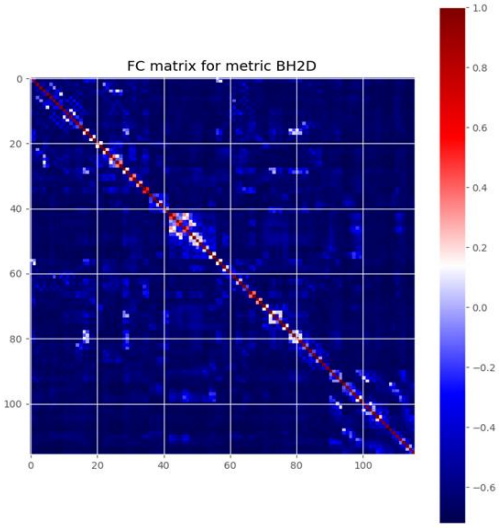

LRP for metric BCohW2

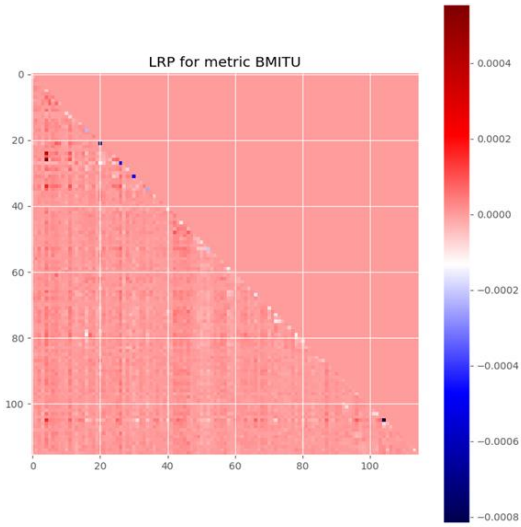

FC matrix for metric BCohW2

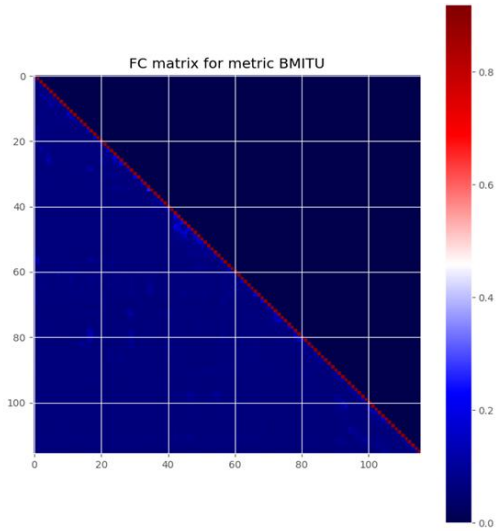

LRP for metric BH2U

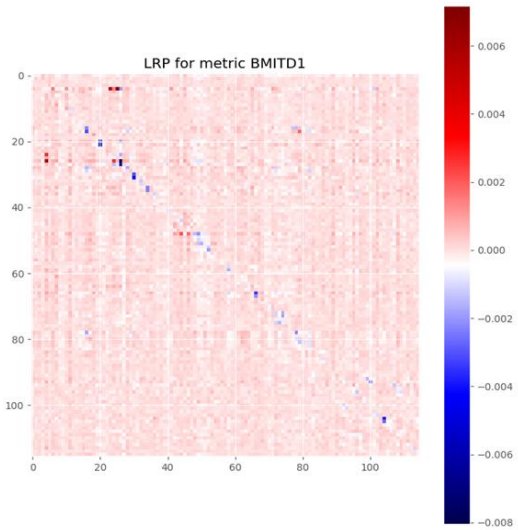

FC matrix for metric BH2U

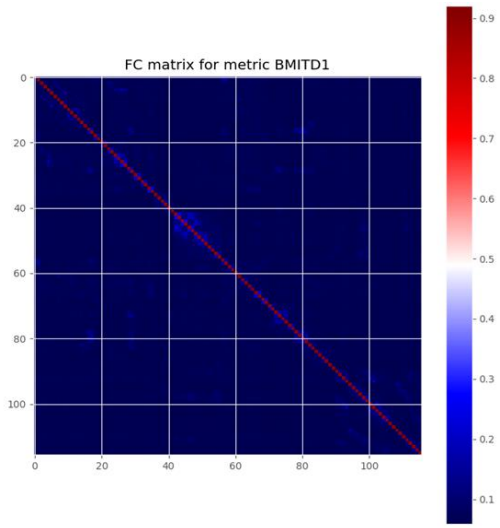

LRP for metric BH2D


FC matrix for metric BH2D


LRP for metric BMITU


FC matrix for metric BMITU


LRP for metric BMITD1


FC matrix for metric BMITD1

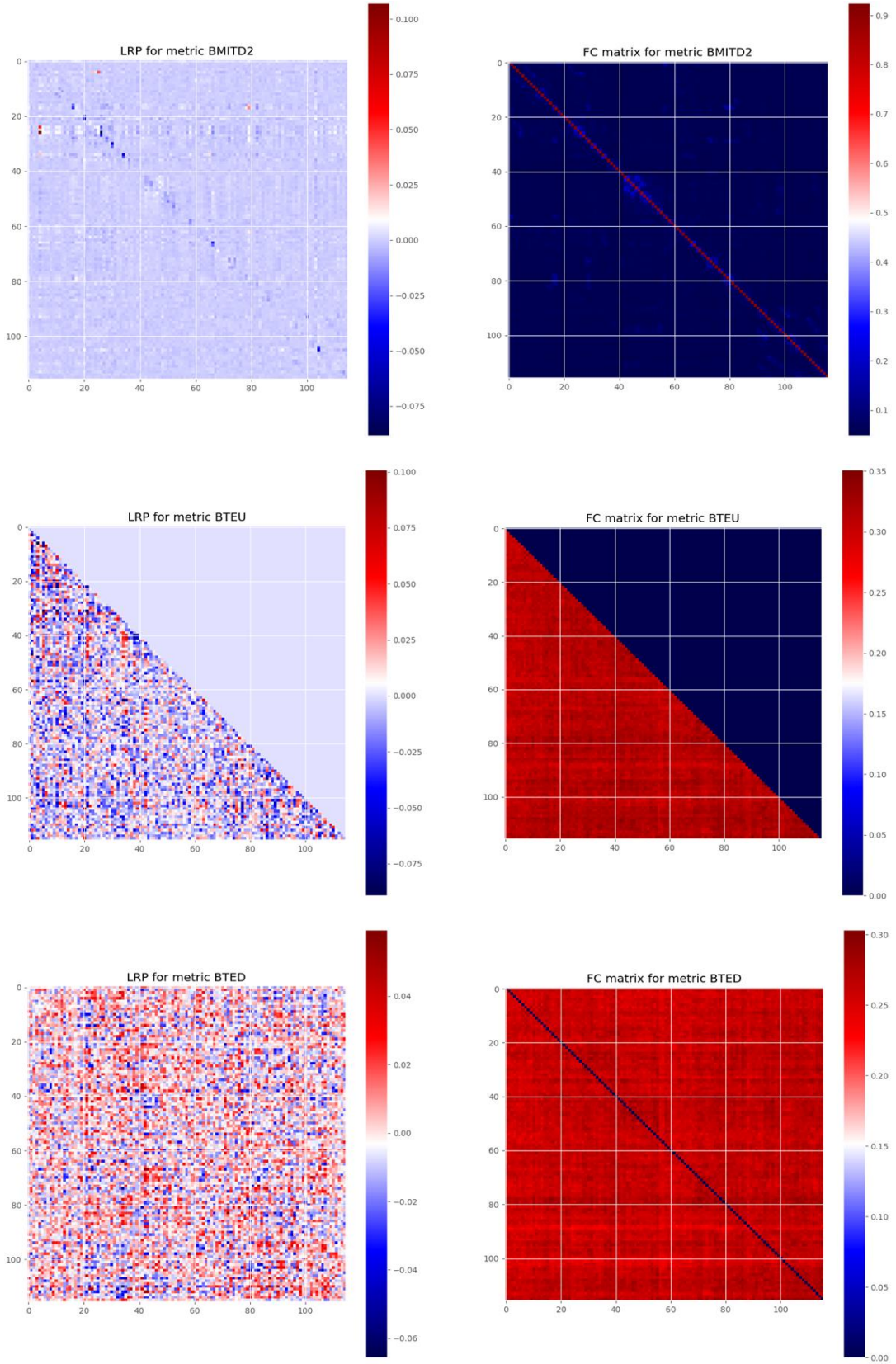
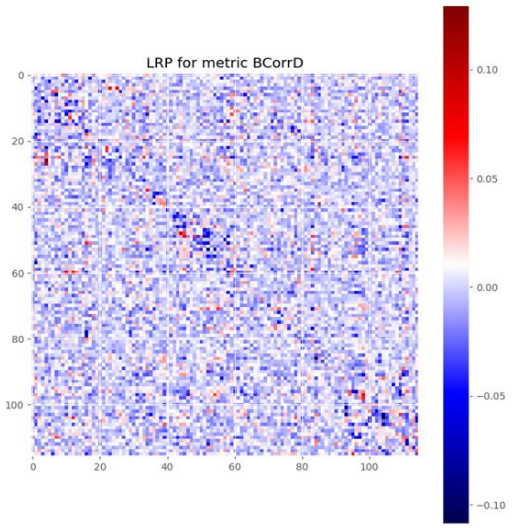
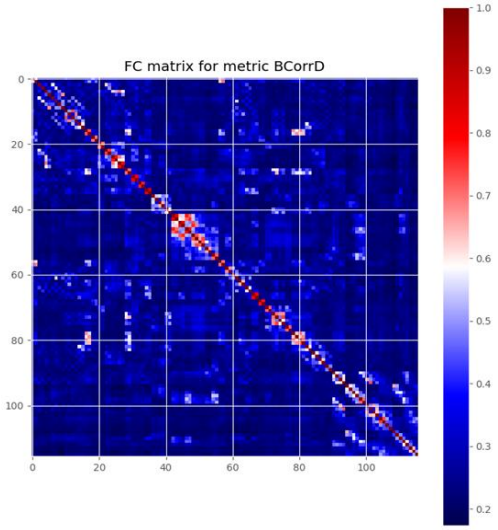
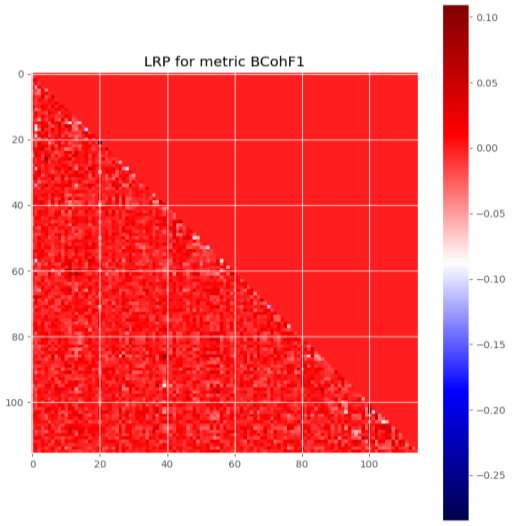
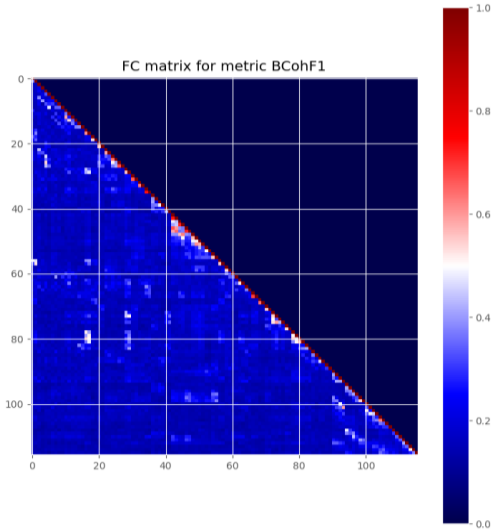


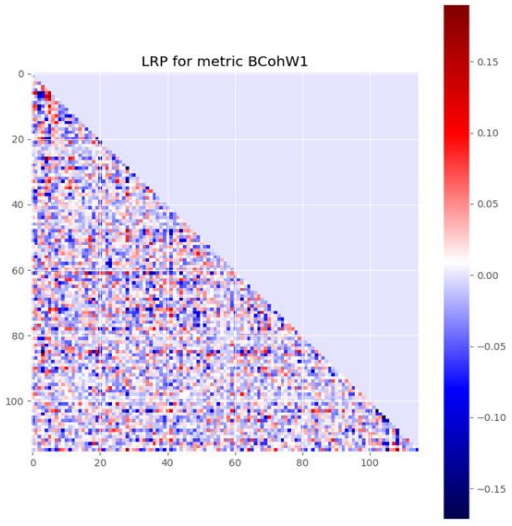
Figure A.2: Heatmaps of the Layer-wise Relevance Propagation analysis for the Functional Connectivity matrices computed with the remaining statistical metrics (left image), BCorrD, BCohF1, BCohW1, BCohF2, BCohW2, BH2U, BH2D, BMITU, BMITD1, BMITD2, BTEU and BTED (top to bottom), and the original FC matrices computed with the respective statistical metrics (right image), when using the ConnectomeCNN model.

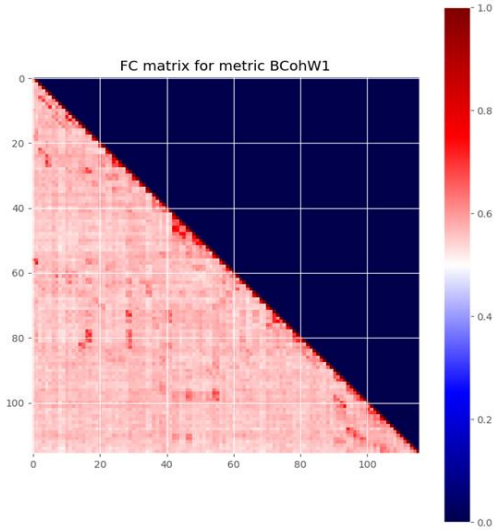

LRP for metric BCorrD

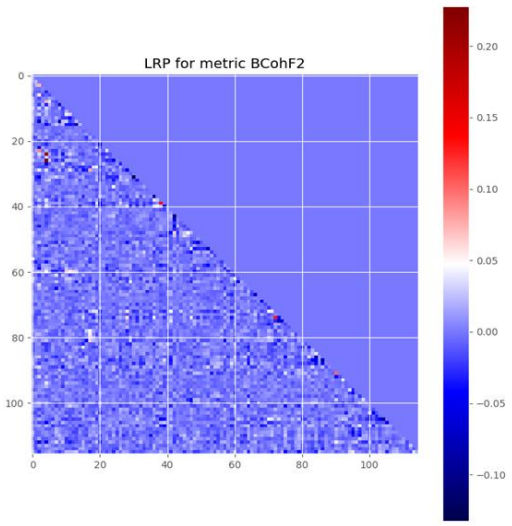

FC matrix for metric BCorrD

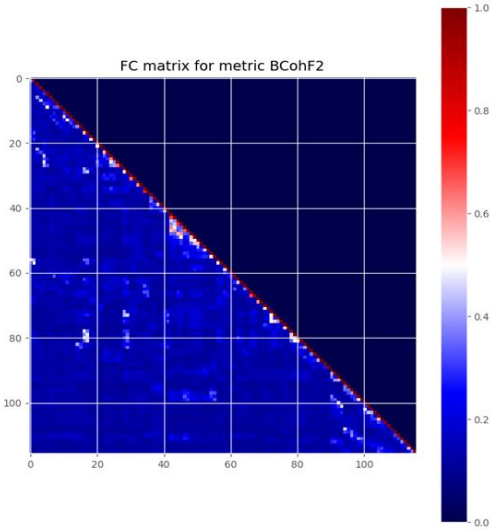

LRP for metric BCohF1

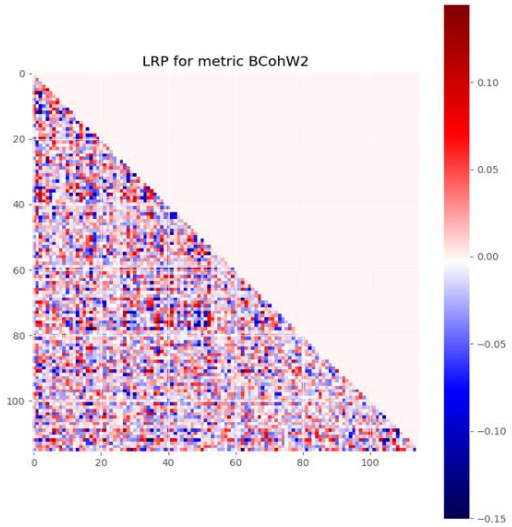

FC matrix for metric BCohF1

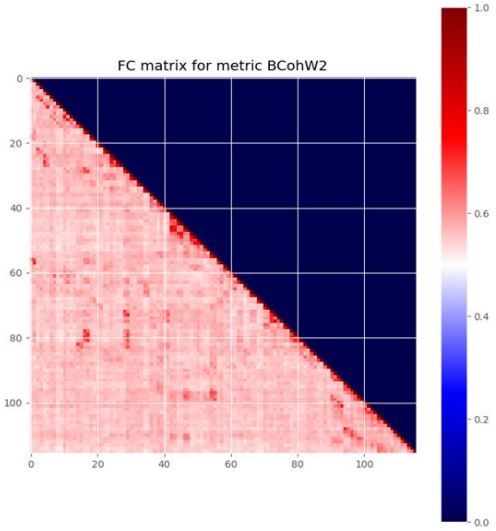

LRP for metric BCohW1

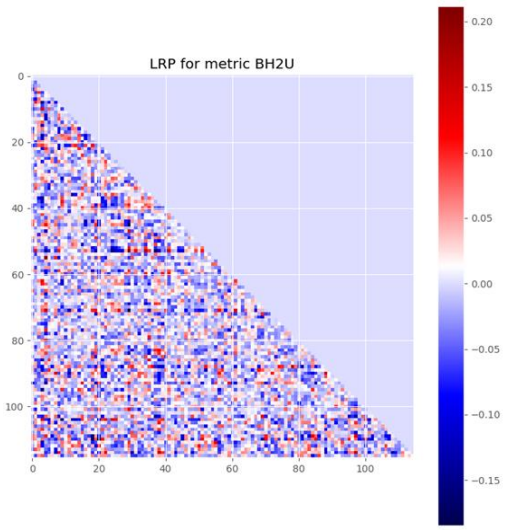

FC matrix for metric BCohW1

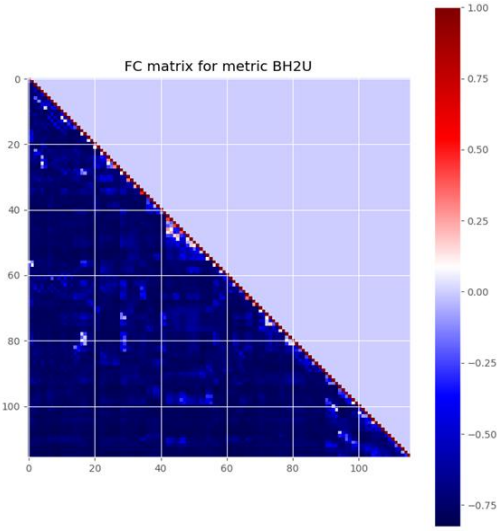

LRP for metric BCohF2

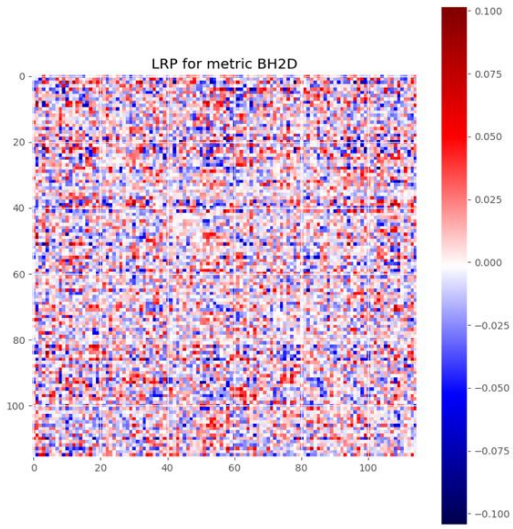

FC matrix for metric BCohF2

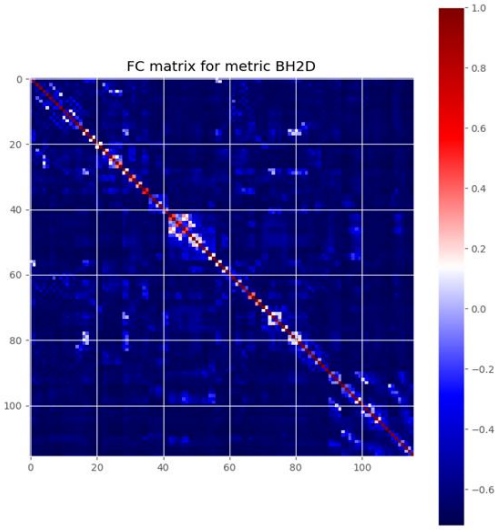

LRP for metric BCohW2

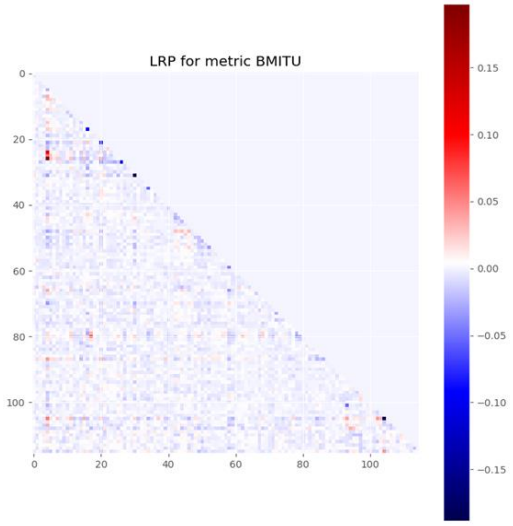

FC matrix for metric BCohW2

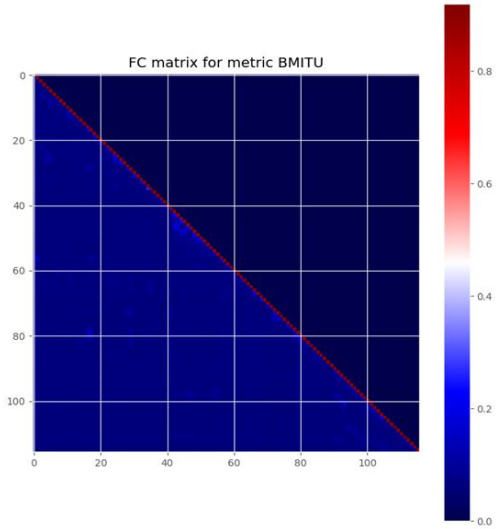

LRP for metric BH2U

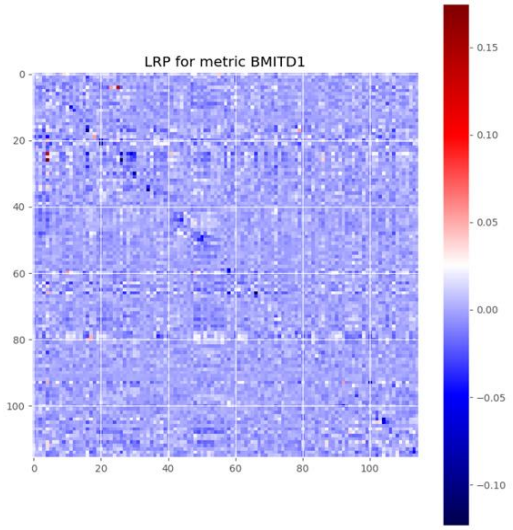

FC matrix for metric BH2U

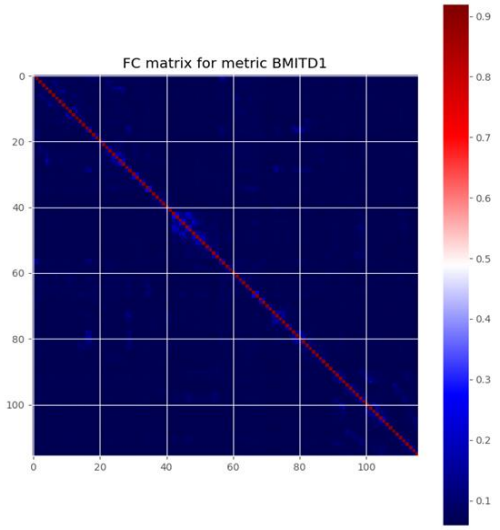

LRP for metric BH2D


FC matrix for metric BH2D


LRP for metric BMITU


FC matrix for metric BMITU


LRP for metric BMITD1


FC matrix for metric BMITD1

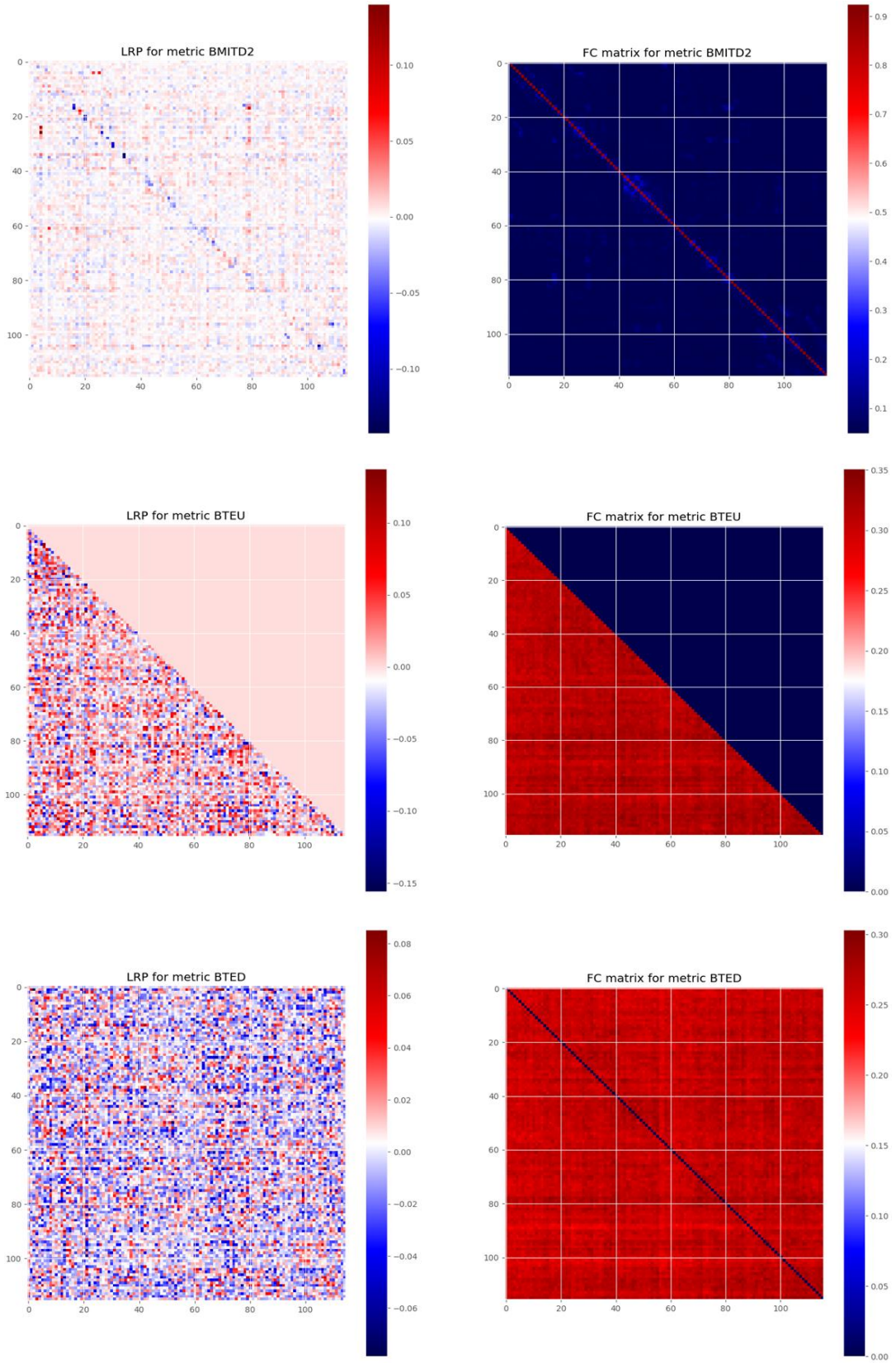


Figure A.3: Heatmaps of the Layer-wise Relevance Propagation analysis for the Functional Connectivity matrices computed with the remaining statistical metrics (left image), BCorrD, BCohF1, BCohW1, BCohF2, BCohW2, BH2U, BH2D, BMITU, BMITD1, BMITD2, BTEU and BTED (top to bottom), and the original FC matrices computed with the respective statistical metrics (right image), when using the ConnectomeCNN-Autoencoder model.