Schizophrenia Diagnosis using Electroencephalographic Data

David António Cóias Calhas

Thesis to obtain the Master of Science Degree in

Information and Computer Engineering

Supervisor: Prof. Rui Miguel Carrasqueiro Henriques

Examination Committee

Chairperson: Prof. Daniel Jorge Viegas Gonçalves
Supervisor: Prof. Rui Miguel Carrasqueiro Henriques
Member of the Committee: Prof. Alexandre Paulo Lourenço Francisco

July 2019
To my family.
Acknowledgments

I would like to thank my family for supporting everything in my life. My supervisors, Rui Henriques and Enrique Romero, for the amazing guidance throughout this last months and accepting me in a critical time. My friends for being my friends. And everyone I met in this incredible journey.
Abstract

The increasing access to brain signal data using electroencephalography creates new opportunities to study electrophysiological brain activity and perform ambulatory diagnoses of neuronal diseases. This work concentrates on applying different approaches in order for one to be capable of making a diagnostic of a neuronal pathology with outstanding confidence. Further, the neuronal pathology taken into account is Schizophrenia, where there is still a research gap when it comes to explore the frequency domain of brain signal data. Hence we fill this gap in two ways: 1) thorough statistical and clustering analysis; and 2) attempt to classify each observation. Given the performance limitations of state-of-the-art classifiers in schizophrenia electroencephalographic data, a new approach was developed. In this context, a pairwise distance learning approach relying on the spectral properties of the signal is proposed. Significant remarks were found using the conducted unsupervised analyzes and the proposed pairwise distance learning approach, with both helping the community on the diagnosis of the Schizophrenia pathology.

Keywords: Electroencephalography, Deep Learning, Hierarchical Clustering, Pairwise Distance Learning, Schizophrenia
Resumo

O crescente acesso a sinais do cérebro com recurso a eletroencefalografia cria novas oportunidades para estudar a atividade cerebral eletrofisiológica e realizar diagnósticos ambulatoriais de doenças neuromusculares. Este trabalho centra-se na aplicação de diferentes abordagens para apoiar o diagnóstico de patologias neurológicas. Neste contexto, este trabalho olha para a esquizofrenia, onde ainda existe uma lacuna quando se trata de explorar o domínio da frequência do sinal electroencefagráfico neste tipo de patologia. Portanto, dadas essas limitações o problema foi solucionado de duas maneiras: 1) análise estatística e técnicas de segmentação; e 2) classificação de indivíduos. De acordo com as limitações observadas na classificação de dados electroencefalográficos em esquizofrenia, propomos uma nova abordagem. Esta abordagem consiste na aprendizagem da distância entre pares para classificação, baseada nas propriedades espectrais do sinal. Os melhores resultados foram encontrados nas análises não supervisionadas e na abordagem proposta para classificação. Estes últimos permitem um melhor diagnóstico da comunidade no diagnóstico de esquizofrenia.

Palavras-chave: Eletroencefalografia, Aprendizagem Profunda, Agrupamento Hierárquico, Aprendizagem à Distância de Pares, Esquizofrenia
Contents

Acknowledgments .................................................................................................................. v
Abstract ................................................................................................................................... vii
Resumo ........................................................................................................................................ ix
List of Tables ................................................................................................................................ xv
List of Figures .......................................................................................................................... xvii
Nomenclature ........................................................................................................................... xix
Glossary ........................................................................................................................................ xx

1 Introduction ......................................................................................................................... 1
  1.1 Hypothesis ....................................................................................................................... 1
  1.2 Thesis Outline .................................................................................................................. 2

2 Background ............................................................................................................................. 3
  2.1 Electroencephalography ................................................................................................. 3
    2.1.1 Frequency bands ...................................................................................................... 3
    2.1.2 EEG properties ....................................................................................................... 3
    2.1.3 Brain regions ......................................................................................................... 4
    2.1.4 EEG channels ...................................................................................................... 4
  2.2 The use of EEG in schizophrenia .................................................................................. 4
  2.3 Processing EEG signal ................................................................................................... 6
    2.3.1 Fourier Transform ............................................................................................... 6
    2.3.2 Short-time Fourier Transform (STFT) ................................................................... 6
  2.4 Distance and Similarity ................................................................................................. 7
    2.4.1 Minkowski Distance ............................................................................................ 7
    2.4.2 Cosine Similarity ............................................................................................... 7
    2.4.3 Dynamic Time Warping (DTW) ........................................................................... 8
  2.5 Clustering ......................................................................................................................... 8
    2.5.1 Hierarchical Clustering ....................................................................................... 8
  2.6 Network Representation and Community Finding ....................................................... 9
    2.6.1 Betweenness Centrality (BC) ............................................................................... 9
    2.6.2 Community .......................................................................................................... 10
5.2 Frequency Features ................................................................. 42
  5.2.1 Hierarchical Clustering ......................................................... 42
  5.2.2 Classification ................................................................. 52
5.3 Siamese Model ................................................................. 53
  5.3.1 Hyperparameter Tuning ....................................................... 53
5.4 Similarity Learned Features .................................................... 53
  5.4.1 Hierarchical Clustering ......................................................... 53
  5.4.2 Classification ................................................................. 57
5.5 Discussion ................................................................. 57

6 Conclusions ................................................................. 59
  6.1 Future Work ................................................................. 60

Bibliography ................................................................. 61
List of Tables

5.1 Leave One Out Cross Validation on FFT frequency extracted features with FFT granularity. 52
5.2 Leave One Out Cross Validation on FFT frequency extracted features with band granularity. 52
5.3 Leave One Out Cross Validation on the Similarity learned features. . . . . . . . . . . . . . 57
# List of Figures

2.1 EEG channel mapping nomenclature (sourced from [6]). ........................................ 5  
2.2 DTW algorithm illustration (sourced from [11]). .................................................. 8  
2.3 Dendogram example. .................................................................................................. 9  
2.4 Illustration of how the sliding kernel convolution operation works (sourced from [19]). . 12  
2.5 On the left, an example of a fully connected network. On the right, the same network with  
   dropout in between layers (sourced from [20]). .......................................................... 13  
2.6 Rectified Linear Unit [21]. ........................................................................................ 14  
2.7 Local optimum versus Global optimum (by Christoph Roser at AllAboutLean.com under  
   the free CC-BY-SA 4.0 license). .................................................................................. 15  
2.8 Confusion Matrix. ..................................................................................................... 17  
3.1 Means and Standard deviation of LZC for each region of interest (sourced from [31]). . . 21  
3.2 Means and Standard deviation of fast rhythms MLZC for each region of interest (sourced  
   from [31]). ................................................................................................................ 21  
3.3 Connectivity Measuring (sourced from [34]). ............................................................... 23  
3.4 Illustration of the Siamese Neural Network proposed in [36] ....................................... 25  
4.1 EEG channel setup. ..................................................................................................... 27  
4.2 Approach pipeline. ..................................................................................................... 28  
4.3 FFT algorithm over a 1 minute EEG channel recording. ............................................. 28  
4.4 Discrete Short-time Fourier Transform of a 1 minute EEG channel recording. ............ 31  
4.5 Architecture of the base neural network. .................................................................... 33  
4.6 Architecture of the neural network used to train the weights for the base network. ......... 33  
5.1 Example of how the distributions were analyzed. ....................................................... 37  
5.2 Percentage difference between Healthy and Schizophrenian individuals, per wave and  
   channel. ...................................................................................................................... 38  
5.3 Euclidean distances between individuals of each set using FFT original granularity. ....... 39  
5.4 Euclidean distances between individuals of each set using frequency band granularity. .. 39  
5.5 Euclidean distances on $he_{and}$ $sez$ and $he_{vs}$ $sez$ sets. ........................................... 40  
5.6 Cosine distances between individuals of each set using FFT original granularity. ......... 40  
5.7 Cosine distances between individuals of each set using frequency band granularity. ....... 40
5.8 Cosine distances on $hc_{and \_scz}$ and $hc_{vs \_scz}$ sets. ................................. 41
5.9 Percentage Difference between each channel using the DTW distance. ................... 41
5.19 Network Analysis, edges weighted with Euclidean Distance. ............................... 42
5.20 Network Analysis, edges weighted with Cosine Distance. ............................... 42
5.10 Dendrogram with all of the individuals (Euclidean Distance). ............................... 43
5.11 Dendrogram with all of the healthy control individuals (Euclidean Distance). ........... 44
5.12 Dendrogram with all of the schizophrenian individuals (Euclidean Distance). .......... 45
5.13 Dendrogram with all of the individuals (Cosine Distance). ............................... 46
5.14 Dendrogram with all of the healthy control individuals (Cosine Distance). ............... 47
5.15 Dendrogram with all of the schizophrenian individuals (Cosine Distance). ............... 48
5.16 Dendrogram with all of the individuals (DTW Distance). ............................... 49
5.17 Dendrogram with all of the healthy control individuals (DTW Distance). ............... 50
5.18 Dendrogram with all of the schizophrenian individuals (DTW Distance). ............... 51
5.21 Network Analysis, edges weighted with DTW Distance. ............................... 52
5.22 Dendrogram with all of the individuals (SNN features with Cosine Distance). .......... 54
5.23 Dendrogram with all of the healthy control individuals (SNN features with Cosine Distance). 55
5.24 Dendrogram with all of the schizophrenian individuals (SNN features with Cosine Distance). 56
5.25 Network Analysis, edges weighted with Cosine Distance on the SNN extracted features. 57
Acronyms

**EEG**  Electroencephalography  
**SNN**  Siamese Neural Network  
**DFT**  Discrete Fourier Transform  
**STFT**  Short-Time Fourier Transform  
**DTW**  Dynamic Time Warping  
**BC**  Betweenness Centrality  
**GN**  Girvan and Newman Algorithm  
**kNN**  K-Nearest Neighbors  
**SVM**  Support Vector Machine  
**NN**  Neural Network  
**MLP**  Multiple Layer Perceptron  
**BO**  Bayesian Optimization  
**REO**  Resting State with Eyes Open  
**CPT**  Visual Continuous Performance Task  
**MLZC**  Multiscale Lempel-Ziv Complexity  
**FL**  Left Frontal  
**F**  Frontal  
**FR**  Right Frontal  
**TLP**  Left Temporal-Parietal  
**C**  Central  
**TPR**  Right Temporal-Parietal  
**PO**  Parietal-Occipital  
**Adaboost**  Adaptative Boost  
**LDA**  Linear Discriminant Analysis  
**TPR**  Right Temporal-Parietal  
**FFT**  Fast Fourier Transform
**DSTFT**  Discrete Short-Time Fourier Transform

**TPR**  Right Temporal-Parietal

**BN**  Base Network

**LOOCV**  Leave-One-Out Cross Validation

**NB**  Naive Bayes

**RF**  Random Forest

**XGB**  XGBoost

**CNN**  Convolutional Neural Network
Chapter 1

Introduction

The recording of increasingly affordable and precise electroencephalographic (EEG) data is creating unprecedented opportunities to understand brain activity, aid personalized prognostics, and promote health through wearable biofeedback systems [1]. Electroencephalography is non-invasive, safe, inexpensive, and shows rich temporal content; in contrast with other brain imaging modalities, such as magnetic resonances, entailing higher costs, risks and restrictions on the periodicity of recordings [2]. EEG monitoring is widely used to assess psychiatric disorders, and has shown to be a valuable source to study Schizophrenia, a disorder affecting about 1% of the world population, largely susceptible to misdiagnoses [3].

Despite the inherent advantages of monitoring electrophysiological brain activity, its use for diagnosing neuronal diseases is still capped by the limited size of case-control populations [4], as well as by the intrinsic difficulties of mining brain signals. Brain signal data is high-dimensional, multivariate, susceptible to noise/artefacts, rich in temporal-spatial-spectral content, and highly-variable between individuals [5].

In this context, with the aim of exploring the rich content of the EEG signal in the frequency domain while tackling the aforementioned challenges, this work proposes two major contributions. First, a comprehensive analysis of the EEG signal: statistical analysis, Hierarchical Clustering (Dendogram, Girvan-Newman Community Finding Algorithm).

Second, a dedicated class of neural networks to extract discriminative features of Schizophrenia from electrophysiological brain data. This approach combines principles from pairwise distance learning and spectral imaging in order to address the aforementioned challenges, enabling superior diagnostics.

1.1 Hypothesis

Problem. A EEG recording or brain signal observation is a multivariate time series $X = \{x^k_j \mid k \in \{1..M\}, j \in \{1..T\}\}$, where $x^k_j$ is a measure of the electrophysiological activity in scalp channel $k$ and instant $j$, $T$ is the number of time points, and $M$ is the multivariate order (number of channels). Given a brain signal dataset $\{(X_i, c_i) \mid i = 1..N\}$, i.e. $N$ EEG recordings $X_i$ annotated with a label $c_i \in \Sigma$, we
aim to identify a discriminative feature space to classify (unlabeled) observations.

The rich properties of the electroencephalographic signal are candidates for the use of deep learning techniques, due to the complexity and relevant content present in it. Although, this is relevant for the study of schizophrenia, their performance in a diagnosis task underachieves, because of the limited number of observations. Subsequently, the hypothesis is that the study in this kind of population and the learning principle in new feature spaces, give us the motivation to develop classifiers with the capacity of achieving better results to diagnose this pathology.

1.2 Thesis Outline

The manuscript is organized as follows. After formalizing the problem, Chapter 2 gives an overview over preliminary concepts necessary to understand this work. Chapter 3 surveys existing contributions on the diagnosis of individuals from brain signal data. Chapter 4 describes the proposed approaches. Chapter 5 shows a thorough analysis of the EEG signal in the frequency domain and how different methods are relevant for diagnosing Schizophrenia. Finally, concluding remarks are drawn in Section 6.
Chapter 2

Background

2.1 Electroencephalography

The brain electrophysiology, measured using electroencephalography (EEG), mainly reflects the summation of excitatory and inhibitory postsynaptic potentials at the dendrites of ensembles of neurons [5]. The electrical fields created by individual neurons aggregate, resulting into effects measurable outside the skull. Contrasting with other brain imaging techniques, EEG signals are known to be cheaper and easier to collect, which make its use very attractive. It provides a excellent resolution in time, having its downside on spatial resolution and a high presence of noise in it. This type of signal is characterized as a time series. A time series $\vec{x}_i \in 0..M$ is a sequence of observations measured along time, where each observation $\vec{x}_i = x_i^0,..,x_i^T$ is characterized by a set of features (a vector in a $T$-dimensional space). It is either univariate if $T = 1$ or multivariate if $T > 1$, and real-valued if features are numeric or symbolic if features are nominal. A tabular dataset is a set of $M$ observations where each observation is a set of features, while a time series dataset is a set of $M$ time series. Given an alphabet $\Sigma$, data can be labeled, $(\vec{x}_i, c_i)$ where $c_i \in \Sigma$.

2.1.1 Frequency bands

Since EEG signals can be seen as a composition of waves produced from different ensembles of neurons, the analysis of their frequency domains are relevant. Statistical factor analysis studies on EEG spectral values yield clusters of frequency components that show considerable overlap with the frequency bands classically accepted, namely infraslow (<0.2 Hz), $\delta$ (from 0.2 to 3.5 Hz), $\theta$ (from 4 to 7.5 Hz), $\alpha$ and $\mu$ (from 8 to 13 Hz), $\beta$ (from 14 to 30 Hz), $\gamma$ (from 30 to 90 Hz), and high-frequency oscillations (HFO; >90 Hz) [5].

2.1.2 EEG properties

[5] provide a good overview of how this signal has contributed to different areas in the field of neuroscience. This work states that certain neuronal oscillations or combinations of neuronal oscillations
measured in EEG signals or combinations of neuronal oscillations are well defined neurophysiological mechanisms that are relevant to understand how cognitive processes emerge. In order to identify an EEG oscillation, one has to show that there is a spectral peak within the frequency band of interest. The oscillation is defined by the peak frequency, bandwidth and power. The authors claim that an assembly of neurons should form a functional entity, which means that a population of neurons of sufficient size should be active in a coordinated way in time and spatially organized, such that their electrical fields may be recordable at distance. There is evidence to believe the brain can be divided into communities (also known as modulating different population sets in the brain) that process information and communicate between each other. Such information and communication processing is key to understand cognitive processes, [5].

2.1.3 Brain regions

The human brain has four main lobes: **Frontal**, **Parietal**, **Temporal** and **Occipital**. The Frontal lobe, located at the front of the brain, is associated with cognitive functions. The Parietal lobe, located at the middle section of the brain, processes sensorial data. It is in this lobe where the somatosensory cortex lies, which again processes the senses. The Temporal lobe, located at the sides of the brain (next to the ears), plays an essential role in the auditory system. Since it contains the primary auditory cortex. This lobe is associated with memories. The Occipital lobe, located at the back of the brain, contains the primary visual cortex and it is known for processing visual stimuli.

2.1.4 EEG channels

The EEG signal is a **multivariate time series**, as it is normally measured using (one or more) electrodes placed at different locations. The electroencephalographic collection at each location is termed channel. The channels are referred to by the region they are located at. For example, in Fig. 2.1 it can be seen each channel nomenclature and its location. The first letters of the EEG identifier provide the lobe location and the following characters identify its specific location in the lobe: $F3, F4, F7, F8, Fp1, Fp2, Fz$ refer to the channels located in the frontal lobe; $C3, C4, Cz$ are located at the mid section of the scalp; $T3, T4, T5, T6$ refer to the channels in the temporal lobes; $P3, P4, Pz$ refer to the channels in the parietal lobe; and $O1, O2$ refer to the channels in the occipital lobe.

2.2 The use of EEG in schizophrenia

EEG imaging has been widely used for the diagnosis of neural pathologies, being a suitable and desirable technique for the schizophrenia pathology. Illustrating, [7] show that the response time was determined to be higher in schizophrenia patients than in healthy controls. Several studies discussed in 3 show that EEG is a technique widely used for the diagnosis and analysis of schizophrenia. A big motivation for the use of EEG to diagnose schizophrenia is its trait of being easily assessed and able to be performed in almost any psychiatric setting as it is shown to be tolerated by almost every patient.
In [8], a status overview on EEG abnormalities is given, in order to diagnose patients with schizophrenia. To this end, they examined the status of development of spectral EEG deviations. In the gathered studies, the meta analysis was limited to those works comparing spectral power between one group of schizophrenia patients and one group of healthy control subjects. The presence of two groups (or populations), one with the pathology and a healthy control group, is essential to identify discriminative features from the gathered signals. The hypothesized differences of schizophrenia individuals were increased delta, increased theta, decreased alpha, and increased beta power. A number of subsequent studies suggested that an increase of activity in the lower spectrum (slow waves) is significantly higher in schizophrenia populations. It is also noted that slow wave abnormality (mainly delta increase) is mainly localized in frontal lobe regions. One of the conclusions [8] is that the delta excess (and to a lesser extent the theta excess) is a strong biological marker of schizophrenia. The suggested EEG locations to be reported when performing schizophrenia diagnosis are: frontal, temporal, central, parietal and occipital electrodes.

Event-related EEG stems suggest that neural oscillations and their synchronization represent important mechanisms for interneuronal communication and binding of information that is processed in distributed brain regions [9]. Several studies with EEGs have in mind the gain of new insights into the pathophysiological processes underlying cognitive deficits in neuropsychiatric disorders, supporting the role of EEG data analysis to study schizophrenia.
2.3 Processing EEG signal

EEG signal is usually characterized as a synchronous multivariate time series. Asynchronous, because it does not show a specific pattern for the whole time series. Multivariate, as there is more than one source of the signal due to the multiple channels present during an EEG recording session. To analyze the EEG signal, a transformation from the time domain to the frequency domain is worth being done, so one can analyze spectral patterns present in it and consequently associate those patterns with brain-evoked events.

2.3.1 Fourier Transform

An important base in signal processing, named after mathematician Joseph Fourier. It decomposes a signal into a set of waves represented by a function of frequencies and intensities. The Fourier transform is usually referred to as the frequency domain representation of a signal. Fourier Transform operates in signals that belong to the continuous world, but because in reality every signal is discrete (being a collection of samples through time) the Fourier Transform cannot be directly applied to discrete signals.

Discrete Fourier Transform (DFT)

The version of the Fourier Transform applied to digital signals (discrete signals). The Discrete Fourier Transform is represented by

$$X_k = \sum_{i=0}^{N-1} x_i e^{-\frac{2\pi j}{N} ki}$$

$k = 0, ..., N - 1$

$X_k$ is the amplitude for the frequency $k$ and there is a total of $N$ frequencies.

Fast Fourier Transform

As the Fourier Transform takes too much computation power, the Fast Fourier Transform was proposed as a faster computation in [10].

2.3.2 Short-time Fourier Transform (STFT)

The STFT divides the signal in windowed segments and computes the Fourier Transform for each of those segments. It is particularly useful when the signal is non-periodic. This comes from the intuition that if you zoom enough on a signal, a periodic segment of a non-periodic signal will eventually be identified. This is the main idea that fits the STFT and also the reason why it is often used in EEG. The window size plays an important role: the bigger it is, the bigger its ability to capture lower frequencies. Particular care will be placed since, as previously said, low frequencies are statistically relevant when distinguishing healthy controls from schizophrenian individuals.
2.4 Distance and Similarity

Given a feature space, and the respective instances of a dataset represented in that space, one can compute distances. These distances can be used to support the analysis of the dataset, namely to facilitate exploration, clustering and classification purposes.

In this section the following metrics will be introduced: Minkowski distance and Euclidean distance; Cosine Similarity; and Dynamic Time Warping distance.

2.4.1 Minkowski Distance

The minkowski distance between two vectors \( a = (a_1, ..., a_n) \) and \( b = (b_1, ..., b_n) \), both in a \( n \) dimensional space, is shown in equation 2.1.

\[
\sum_{i=0}^{n} |a_i^p - b_i^p|^{\frac{1}{p}}
\]  

(2.1)

Euclidean Distance

The euclidean distance between two vectors \( a = (a_1, ..., a_n) \) and \( b = (b_1, ..., b_n) \), both in a \( n \) dimensional space, is shown in equation 2.2.

\[
\sqrt{\sum_{i=0}^{n} a_i^2 - b_i^2}
\]  

(2.2)

It is a specialization of the Minkowski distance (equation 2.1) with \( p=2 \).

2.4.2 Cosine Similarity

Cosine Similarity can be seen as the correlation between two vectors. It is represented by the formula in equation 2.3. Where \( A \) and \( B \) are instances.

\[
\cos A, B = \frac{A \cdot B}{\sqrt{A^2 \cdot B^2}}
\]  

(2.3)

The cosine function takes values in the interval \([-1, 1]\). Being \(-1\) when two vectors take opposite directions (180° angle), \(0\) when two vectors are perpendicular to each other (90° angle) and \(1\) when two vectors have the same directions (0° angle).

Just as the cosine function can be seen as a similarity measure, one can also derive distance value out of it, which is known as the Cosine Distance. Cosine Distance can be computed by equation 2.4, taking values in \([0, 2]\).

\[
d = 1 - \cos A, B = 1 - \frac{A \cdot B}{\sqrt{A^2 \cdot B^2}}
\]  

(2.4)
2.4.3 Dynamic Time Warping (DTW)

The introduced distances were originally proposed to compare vectors in n-dimensional spaces. Although they can be applied to time series, they disregard the temporal dependence between values and therefore are unable to adequately account for temporal misalignments.

As the EEG is a time series, the DTW is a suitable algorithm to compare two EEG recordings. Consider for example, that given a feature from the EEG that varies along time, then one can use the DTW algorithm to compute a distance between two EEG instances, just based on that feature.

![DTW algorithm illustration](sourced from [11]).

What the DTW algorithm does is: given two time series, it performs an optimal match between them. This match is done by warping two signals and computing the distance between them. The warping with the smallest distance computed will be chosen and it will be the optimal match.

2.5 Clustering

When working with a dataset that has more than one class of samples, one can hypothesize and explore whether data is well separated using its features and different groups of the same class having coherent distributions and whether the observations from the same class have coherent distributions. Clustering is the field of study that concentrates on the grouping of observations of a dataset into different clusters.

2.5.1 Hierarchical Clustering

Among the existing types of clustering, this thesis will primarily consider Hierarchical Clustering. This type of clustering technique identifies relations between different groups of instances and how those groups are related.

A Dendogram is a useful means to visualize the organization of instances in a dataset. An example of a dendogram is presented in Fig. 2.3. Given a distance matrix, observations with a close distance are grouped in the same group, and far distanced instances being placed far apart in the dendogram, as the clusters grow connections between them are also made taking into account the distances between the instances of each cluster. A good use for dendograms is the ability to cut it, according to a threshold, and through this cut having different groups (clusters) of data.
2.6 Network Representation and Community Finding

Given a set of instances and a distance matrix, one can model a network. A network/graph is represented by $<V, E>$, where $V$ is the set of nodes and $E$ the set of the edges connecting the nodes. The modelling can be done with the instances being the nodes and the edges connecting them being based on a relation. This relation can be thought of either considering a threshold, $t$, by which instances with distances lower/higher than $t$ will have a link connecting them (unweighted undirected graph). Another way of modelling a network based on a distance matrix is by representing each edge with weight equal to the distance between the instances it is connecting (fully-connected weighted network). Mixing these representations based on distance can also be useful in order to avoid a high density of edges, where there exists a threshold $t$ and each edge has a weight. The following concepts – Betweenness Centrality, Community and Girvan and Newman Algorithm – are going to be explained, with the motivation of being used to perform a clustering analysis.

2.6.1 Betweenness Centrality (BC)

A common topological feature to study the properties of a network is the BC measure. It assigns to each node/edge a value, and the bigger the value the bigger its BC. Nodes/edges with high BC often play a more important role on the network, as they are responsible for a high flow of information.

To check which nodes/edges have the highest BC, one can compute all the shortest paths between every pair of nodes and check which nodes/edges are present in most of those paths. A shortest path between a pair of nodes is a path such that the number of edges present in it is lower than all of the other paths between that pair (for unweighted graphs), or the sum of the weights of the edges is the lowest (for weighted graphs).
2.6.2 Community

Given a network, a community is a subset of coherently connected nodes, commonly densely connected. In order to identify communities in a network, a search on several partitions is made in order to assess modularity, 2.5, where $\frac{E_{r|\text{net}}}{E}$ is the number of internal links in module $r$ and $(\frac{k_r}{2E})^2$ is the number of internal links in module $r$ for a randomized network. $k_r$ is the sum of the degrees of the edges of module $r$. $E$ is the total number of edges of the network [12].

$$M = \sum_{r=1}^{n} \frac{E_{r|\text{net}}}{E} - (\frac{k_r}{2E})^2$$  \hspace{1cm} (2.5)

The modularity variant defined by equation 2.5 is applied for non weighted graphs. For weighted graphs, an additional variant is applied as described in equation 2.6, where $W_{r|\text{net}}$ is the sum of the weights of the internal edges of module $r$ and $S_r$ is the sum of the strength of the vertices of module $r$. The degrees are replaced by the sum of the weights of edges adjacent to the vertex (strength), and the number of edges by the sum of weights (W). [12].

$$M = \sum_{r=1}^{n} \frac{W_{r|\text{net}}}{W} - (\frac{S_r}{2W})^2$$  \hspace{1cm} (2.6)

2.6.3 Girvan and Newman Algorithm (GN)

The community finding problem is a computationally expensive task and there is a wide number of algorithms to perform it. The GN algorithm [13] is iterative and progressively removes the nodes/edges with highest BC, because removing these type of nodes are the most probable to be connecting communities. The remaining connected partitions of the network, after each node/edge is removed, are the communities. The nodes/edges are removed until none remains. The output of the algorithm is a dendogram and, using this dendogram, one can perform hierarchical clustering based on a network that contains the relations between the nodes. GN is useful when one knows apriori the number of partitions of the network.

2.7 Classification

The main goal of this thesis is to perform a classification task on brain signal data for the diagnosis of the schizophrenia pathology. Given a labeled dataset, the classification task aims to learn a mapping model between data instances and corresponding classes in order to accurately label new (unlabeled) instances. In this section, some classification algorithms are going to be introduced, namely: K-Nearest Neighbors, Support Vector Machines and Neural Networks.

2.7.1 K-Nearest Neighbors

There are several classification algorithms that operate based on the distance between instances. One example is the K-Nearest Neighbor (kNN) algorithm [14]. The kNN algorithm classifies an instance
based on the $K$ nearest instances. Given a distance matrix the algorithm will check what are the $K$ instances with the smallest distance to the instance under classification, and will make its prediction based on their labels. With this kind of algorithm, one just has to make sure that the feature space where the algorithm operates offers a robust way to measure distances.

### 2.7.2 Support Vector Machine

A SVM [15] can be either used for regression or classification tasks, being considered for the second task in this thesis. This algorithm finds a separation between a set of points $S = x_1, ..., x_N$, with $x_i \in \mathbb{R}^m$ for any $i \in [1, N]$. This separation is performed via drawing of a decision line (hyperplane) that maximize separation between the data points from different classes, and also fixing support vectors defining a margin. These support vectors are fixed taking into account the closest samples to the decision line. This is a proper behavior when the data points are linearly separable.

Now take the following example, where a set of points $S' = x'_1, ..., x'_N$, with $x'_i \in \mathbb{R}^m$ for any $i \in [1, N]$, is not separable. In this case, an SVM performs a transformation/mapping, $\Phi(x'_i)$, from space $\mathbb{R}^m$ to a space where the mappings of the points in $S'$ are closer to be linearly separable. This mapping is also known as the kernel trick, where before any labelling of data is done, the data is first mapped to a new space. Given those points and a linear decision boundary (the hyperplane), each one of them is given a label depending on which side of a decision line it is located.

### 2.7.3 Neural Networks (NN)

A NN [16] consists on a collection set of neurons connected between each other. Each connection has a weight value, and each neuron has an activation function. During the learning session, the weight values are corrected, in order to minimize the error between the output of the NN and the target value. This correction of weights can be done using the backpropagation algorithm.

NNs are used as state of the art techniques in multiple fields. A description of them is provided in section 2.8.

### 2.8 Deep Learning

Nowadays, Deep Learning techniques [17] are known to have outperformed traditional models in areas like: classification, generation, regression, transformation, segmentation and feature extraction. In this section, feed forward neural networks, different types of layers, activation functions, loss functions and optimization functions are going to be introduced.

#### 2.8.1 Feed Forward Networks or Multiple Layer Perceptron (MLP)

A MLP is the simplest type of NN, apart from a single neuron network. The input of this type of NN is $x \in \mathbb{R}^n$, where $n$ is the number of features in the input. The goal of a MLP is to learn a non-linear
function $f^*$ that maps input data observations into a desirable output [18], $\hat{y} = f^*(x)$ such that $y = \hat{y}$, where $y$ is the desired output. This type of network is described by a sequence of layers. Each layer $f^i$ is characterized by a set of neurons, with each neuron being a linear composition of values obtained from previous layer or input. In other words, the activations from the previous layer serve as input for the next layer. Until the values reach the output. One can see a MLP as a chain of functions $F$

$$F(x) = f^l(f^{l-1}(...(f^1(x))))$$

where $l$ is the number of layers in a network and $x$ is the vector of features at the input. $n$ can also be seen as the length/depth of the model. When the depth is high these type of learning techniques are referred to as Deep Learning [18]. Each layer has a collection of neurons, which compute calculations on the inputs they receive (outputs of the previous layer) with an activation function, $a$.

$$f^k = a(\sum_i^n w_j x_j + b)$$

where $f^k$ is the $k$th neuron of a layer, $a$ is the activation function that receives $N$ inputs, $w_j$ is the weight of the connection from the $j$th neuron of the previous layer to this neuron, and $x_j$ is the output of the neuron in the previous layer. These $w_j$ and $b$ are the values to be optimized during a learning session.

### 2.8.2 2-Dimensional Convolutional Layers

In this thesis, a Convolutional layer or a convolution is referred to as an operation performed specifically in the 2-dimensional space of the form $\mathbb{R}^{nm}$, with $n > 0$ and $m > 0$. These type of layers receive as input a vector $x \in \mathbb{R}^{nm}$. Based on the nomenclature in [18], one can define a convolutional layer as

$$Y = I \ast K$$

where $I$ is the input, $K$ is a selected **kernel**, and $Y$ the output, sometimes referred to as **feature map**. The $\ast$ symbol represents the convolution operation. A single-step convolution is illustrated in Fig. 2.4.

![Figure 2.4: Illustration of how the sliding kernel convolution operation works (sourced from [19]).](image)
2.8.3 Dropout Layers

This layer was first proposed in [20]. Its purpose is to drop a certain amount of connections between layers, so that the connections maintained are able to learn and perform without the help of other connections. Specifically, for each batch given to the network during the training session, connections between layers have a probability $p$ of being dropped, which means they will have no impact on the output of the network and will not have an update during the backpropagation. With this technique the dependence of the network in a small set of connections in between layers is avoided, subsequently giving a better performance of the model in a test set.

Figure 2.5: On the left, an example of a fully connected network. On the right, the same network with dropout in between layers (sourced from [20]).

2.8.4 Activation Functions

In machine learning, and specifically in NN, an activation is the output of an activation function of a neuron. It comes from the biological process of a firing neuron going through the axon and reaching the other neurons dendrites at the synapses. The same is emulated in neural networks when a neuron "fires" its output from the activation function, and that output goes through the edge connection (where a weight is associated, representing the inhibition or excitation of an activation). The value from the activation is multiplied by the weight of that edge connection. There is a multitude of activation functions. One of those functions is the Rectified Linear Unit, first introduced in [21], it is a function that takes values $\geq 0$, being defined by $\max(0, x)$ as seen in Fig. 2.6.

Another important activation function is the Softmax. This function is particularly useful as its output vector always sums to 1. Its definition can be seen in equation 2.7, where $j = 1, ..., n$ and $x = (x_1, ..., x_n)$.

$$
\sigma(x)_j = \frac{e^{x_j}}{\sum_{i=1}^{n} e^{x_i}} \tag{2.7}
$$
2.8.5 Loss and Optimization functions

In order for a neural network to learn, it needs objective/quantitative guidance. This guidance is provided by the Loss and Optimization functions.

Loss functions

The role of these functions is to compute a difference between the output of the network (predicted value) and the true value of the instance given as input (target value).

In particular, Contrastive Loss, which is of interest to this work, is a loss function that works with pairs of samples. Introduced in [22], it computes distance between pairs of instances, such that neighbors are pulled together and non-neighbors are pushed apart, according to a certain distance metric. The Contrastive Loss function is defined by:

$$L(W, Y, X_1, X_2) = Y D_W^2 + (1 - Y)\max(0, m - D_W)^2,$$

with $L$ being the function; $Y$ the true values of the distances between the pairs; $D_W$ the predicted distance values; and $m$ being the margin factor of separation. The margin value can be tweaked, and too high values can make the model training session longer (but on the bright side it increases the separation between different pairs, as the margin is bigger) and too low values may cause the model to not learn at all. [23] found that higher values to have a positive impact on the results of its model.

Optimization functions

The guidance during the training session of a neural network is done specifically by this type of functions. There are many studies concerning these functions as they determine the ability to learn. The main goal is for the set of weights of the whole network to have a global optimum value, such that the network always predicts the right value for every possible instance. In other words, a set of weights that makes $f^*$ (approximated function, learned by the network) behavior as close as possible to $f$ (real function) behavior. Although reaching a global optimum for this task is from impossible to extremely unlikely, one can reach a local optimum set of weights that performs optimally or close to optimal for a set of instances. And this is where the field of research on optimization comes in.
The learning rate is one of the most important parameters of an optimizer. A high learning rate may cause the model to miss a global optimum, and in contrast a low one makes the training session longer as it spends more time reaching the global optimum, if it reaches that optimum at all. Another difference between high and low learning rates, is that low learning rates give a smoother loss decrease during the training session avoiding sudden increases/decreases in it, making the learning session more stable.

Momentum accumulates the gradient of the last updates (just as velocity) and has influence in the updates of the next steps (direction of the gradient). It is very useful and optimizers that implement it usually have very interesting convergence results.

One state of the art optimization function is RMSprop, it stands for Root Mean Squared propagation. First introduced in [24], it is now a widely used function. It is known to adjust the learning rate for each parameter (weight). Another function, which can perform better than RMSprop depending on the type of architecture of the model, is the Adam optimization function, unveiled in [25]. This function follows principles from RMSprop and also implements momentum.

2.8.6 Regularization

Optimization functions like RMSprop and Adam are known to reach a local optima really fast, which sometimes may not be advantageous for the learning. Also, these optimization functions produce a neural network whose outputs rely heavily on a small set of weights, leaving other weights with low importance in its final output. To tackle this problem, regularization methods are widely used to train neural networks. They are known to avoid spontaneous big changes in loss values during the training session, helping the model to reach better results in test sets and giving a smoother decrease of loss values during each training session. There are two major types of regularization methods used: L1 and L2 regularization.

L1 regularization

Regularization can be applied to the kernel (weights), bias or activation values. Regularization in the kernel and bias values is done during their updating, through backpropagation. The intuition is the addition of a new term, the regularization expression \(-\lambda \cdot ||W||_1\), being \(\lambda\) the regularization constant and
\( W \) the weight matrix. For the bias and activation regularization, the \( W \) term just has to be replaced by the bias and activation values.

**L2 regularization**

Just as L1 regularization, L2 regularization can be applied to kernel, bias and activation values. Its expression is defined as \(-\lambda \cdot ||W||_{L2}\), being \( \lambda \) the regularization constant and \( W \) the weight matrix. The main difference between L1 and L2 regularization is that L1 regularization can more easily approach 0. This property makes L1 more suitable in the presence of uninformative features as it performs feature selection itself.

### 2.8.7 Hyperparameters

In deep learning, hyperparameters are parameters of a network that can not be learned. The only parameters that can be learned in a neural network are the weights and biases. Parameters like layers, type of layers, activation functions, loss functions, among others, are referred to as hyperparameters.

Although, optimizing the network is essential for it to learn, there is a second optimization process that is done when building a NN. Hyperparameter tuning is, as its name points out, the tuning of the hyperparameters. This is often made manually, which can take a long time, with a lot of people performing a standard grid search (also very expensive computationally) where one defines different domains for different hyperparameters. [26] introduce a hyperparameter tuning process using Bayesian Optimization, which they claim is much faster than a standard grid search and has shown to reach better results than other optimization techniques as well as human manual tuning.

The BO algorithm [27] does not take into account the type of function it is optimizing. Instead, there is an initial exploration made, which defines a prior distribution, in order to capture initial beliefs about the function. Following this initial exploration, another exploration is done updating the prior distribution to a posterior distribution. This posterior distribution is built using an acquisition function, which makes the decision of each point should be explored next.

### 2.9 Model Validation

Once a model is built, the validation step is performed. In order to validate a model, a partition of the dataset – train set, validation set and test set – is done. The train set is the partition of the original dataset one uses to learn the model. The validation set is used so one can know the progression of the model during its training session. The test set is used to see how our model behaves with never seen before data. Given the properties of a dataset, the use of these three partitions (instead of simple train-test partitions) are extremely important when developing and validating a deep learning model.
2.9.1 Cross Validation

Usually, a model should not be validated in just one partition of train, validation and test sets, since specific partitions can benefit the performance of the model than others. To avoid this, Cross Validation comes to the rescue by creating different partitions and performing validations in each one.

K-fold Cross Validation

In this version, a fixed size proportion of the dataset is allocated to train and test set. The model is validated $K$ times in $K$ different partitions of train and test sets.

Leave-one-out Cross Validation

Usually, when working with small datasets, K-fold cross validation with $k << N$ can hamper the learning by undesirably reducing the size of the train set. Leave-one-out performs $n$ validations, each learning a model from $n - 1$ training instances and testing it on the remaining (test) instance.

2.9.2 Metrics

Given a confusion matrix, which can be seen in Fig. 2.8, one can compute the following evaluation metrics:

- Accuracy is defined as $\frac{TP + TN}{TP + TN + FP + FN}$
- Sensitivity is defined as $\frac{TP}{TP + FP}$
- Specificity is defined as $\frac{TN}{TN + FP}$

![Confusion Matrix](image)

Figure 2.8: Confusion Matrix.

In order to determine the significance of the results (accuracy, sensitivity and specificity), the p-value is computed. This is based on an hypothesis statistic test, specifically against the null hypothesis.
Chapter 3

Related Work

3.1 Schizophrenia Statistical Analysis

It is known that increased delta activity and decreased alpha activity, apparent in psychosis, are indicative of unusual/inappropriate arousal states, which leads to reduced ability to attend to relevant information. In [28], it is also noted that delta/alpha frequency activity may provide a useful neurophysiological biomarker to identify psychotic disorders. The use of EEG is justified in [28] since it offers a reliable measurement of direct brain activity with higher temporal resolution than functional imaging. The EEG patterning of neuropathologies, such as schizophrenia, bipolar disorder and methamphetamine-induced psychotic disorder are studied in this work. A healthy control group is also considered to differentiate populations. The dataset consists of three types of EEG recording performed by trained clinical personnel: resting state with eyes open (REO), resting state with eyes closed (REC) and a visual continuous performance task (CPT). EEG records of 3 minutes were obtained for each of the resting state conditions. The recording was undertaken using a simple EEG montage that included frontal (F3 and F4), central (C3 and C4), and parietal (P3 and P4) electrodes. 109 research participants participated in the study: 28 individuals with diagnosis of Schizophrenia, 28 with a diagnosis of Bipolar disorder, 24 with diagnosis of Methamphetamine-induced disorder and 29 healthy control individuals. All participants were aged between 19-40 years old. For the analysis between groups, first Kruskal-Wallis multivariate analysis of variances is performed, as well as correlation studies using Spearman’s rank order correlation applied across and within groups (in accordance with their respective delta/alpha activity). The main finding was that a higher delta and lower alpha synchronization during the three testing conditions is able to delineate psychotic disorders, such as schizophrenia. Delta synchronization over frontal cortex has been associated with poor cognitive performance, and it was apparent for the schizophrenic individuals. However, relative delta/alpha activity was not correlated with behavioural performance in the current study. No relationships with age or duration of illness were found for delta/alpha frequency activity, suggesting age and duration of illness were not potential confounders. This work gives an overview how psychotic disorders can produce different brain signal patterns when both comparing the different disorders, as well as healthy patients and patients with one of the pathologies analyzed. It motivates this
thesis and the use of the EEG technique for schizophrenia diagnosis.

Individuals that suffer from schizophrenia exhibit visual processing deficits, namely, the ability to form global objects from local elements. In [29], a study on the visual integration neural process is made through EEGs, leading to observable/measurable differences in individuals with schizophrenia and healthy controls. A dataset was gathered, where 48 individuals with schizophrenia and 31 healthy controls participated. These participants viewed images that contained illusory, non illusory and real contours. Time of response was measured, as well as frequency analysis between schizophrenia individuals and healthy controls. The results show that dysfunctional neural synchronization is associated with visual integration in schizophrenia. Schizophrenia patients failed to show phase locking in the early (< 100 ms) 40-60 Hz gamma band response to illusory stimuli, suggesting that poor neural synchronization during sensory processing may underlie visual integration deficits. And according to [29], since the ability to organize thoughts and behaviour may share the same neural process as visual integration, it may be possible that when one process is dysfunctional it leads to dysfunction in the other. This work gives further insight on the schizophrenia pathology with an analysis of a visual task.

In [30], non linear relation between EEG signals and schizophrenia diagnosis was tested. The dataset used had 18 healthy controls and 18 schizophrenia patients. The EEG data was recorded from 14 electrodes in total, with a sample rate of 100Hz. To test the non linearity hypothesis, a correlation dimension with a spatial embedding was applied, and consequently used as a discriminating statistic. One-way Analysis of variance was performed to analyze the non-linear measure. A marked decrease in dimension complexity was found in schizophrenia EEG signals compared with controls using spatial embedding methods. This decrease in dimensionality of schizophrenic subjects was due to an hypothesized decrease in complexity of the brain electrical activity. Also, decreased dimension complexity reflects that slow wave activities have increased, which suggests overall brain dysfunction. This work motivates the use of non linear methods to diagnose schizophrenia through EEG signals.

Since EEG signals are the result of the non-linear combination of electrical activity generated by interacting oscillators from the cerebral cortex and other biological sources, complex non-linear structures rise when looking at the EEGs in the time dimension, [31]. Consequently, non-linear measures might be more adequate (than linear approaches) to reflect the complex, irregular and non-stationary behaviour of neural processes. The goal in [31] is to gain further insight into how brain complexity changes under cognitive demands in patients with schizophrenia and healthy controls. The dataset recorded consisted of 18 subjects diagnosed with schizophrenia and 17 healthy control individuals. All the participants were shown 80 simple black and white pictures and were asked to name the pictures. During this task, each individual had the brain signal recorded through EEG. The method chosen to analyze the EEGs was the Multiscale Lempel-Ziv complexity (MLZC) and its selection relies on the advantages over other non-linear measures [31]. MLZC can be applied to short time series and non-stationary and noisy signals. Lempel-Ziv Complexity measures complexity, namely, the number of bits of the shortest computer program that can generate the analyzed time series. The original Lempel-Ziv Complexity measure estimates the complexity of a time series by a binarization process in which the signal is transformed into a binary sequence by using its median as a threshold. MLZC can be seen as a generalization of the
original method, because it uses multiple thresholds for binarization. Results from [31] can be seen in figures 3.1 and 3.2. The regions of interest were: Left frontal (FL), Frontal (F), Right frontal (FR), Left temporal-parietal (TPL), Central (C), Right temporal-parietal (TPR), Parietal-occipital (PO).

Figure 3.1: Means and Standard deviation of LZC for each region of interest (sourced from [31]).

Figure 3.2: Means and Standard deviation of fast rhythms MLZC for each region of interest (sourced from [31]).

Results show that patients with schizophrenia exhibited higher complexity in frontal regions than control participants at rest. From this work, an interesting feature for schizophrenia diagnosis was introduced: MLZC, a complexity measure of the EEG signal. According to [31], EEG complexity is being increasingly used to explore brain dynamics in healthy and pathological states. This is due to complexity indexes might more adequately reflect the complex, irregular, non-stationary behaviour of neural processes more than traditional event related potential features.

According to [32], endophenotypes are alternative phenotypes reflecting internal phenomena of organisms that ideally define elements of mental disorders proximal to effects of genes that cause psychopathology. To perform the analyses and take conclusions, in [32], the authors analyzed 48 schizophrenia patients, 61 first-degree biological relatives of schizophrenia patients, 30 bipolar disorder patients, 38 first-degree biological relatives of bipolar relatives of bipolar disorder and 79 demographically similar control participants. Two types of recordings were done: one with the eyes closed, and with
the participants sat down and relaxed; and another with the eyes open, where the participants also sat down relaxed while viewing a dot to keep their eyes fixated. Both sessions lasted 3 minutes. Differences in both conditions were analyzed with multivariate analyses of covariance. Non parametric Wilcoxon tests were used to assess the scalp topography of simple effects, and Pearson correlations were used to assess associations between EEG variables. Regarding the spatial distribution of standard EEG frequencies in both resting conditions, slow beta was maximal across central-parietal scalp sites. Fast beta was predominately distributed across central sites in the eyes-closed condition, but in the eyes-open condition it was most prominent at frontal-temporal scalp sites. Gamma was primarily distributed across frontotemporal scalp sites. Also, the frequency components were highly similar across eyes-closed and eyes-open conditions. As conclusions, elevated delta power differentiated schizophrenia patients in the population taken into account. And evidence suggests that high-frequency activity in resting state EEGs may function as an endophenotype for schizophrenia. Another important conclusion taken in [32], was that excessive EEG high frequencies in schizophrenia may relate to anomalies of the prefrontal cortex that are neural manifestations of genetic liability for the disorder. This work, remains consistent with other studies regarding schizophrenia and EEG.

3.2 Schizophrenia Classification

According to [33], mostly changes in functional connectivity are seen in schizophrenia patients, as well as differences in theta-frequency activity. This motivated the work done in [33] to perform classification of schizophrenia patients. This classification approach involved one single electrode, with a signal of 1 minute duration. With this, differentiation between healthy subjects and subjects who have been diagnosed with schizophrenia and were receiving relevant prescribed medication. An EEG dataset containing a group of 50 individuals (25 schizophrenia patients and 25 healthy controls) was collected. A stimuli was given during the recording session, where black triangles on a gray background were presented in a predictive sequence. The predictive sequence consisted of three types of triangles facing left, up and right, always in that order. The developed system is called Time-Frequency transformation followed by Feature-Optimization. It consists of four stages: preprocessing and the breaking of the signals into relevant intervals; transformation of the EEG signal into a time/frequency representation via the Stockwell transformation; feature extraction from the time/frequency representation; and discrimination of specific time frames following a given set of stimuli between the time/frequency matrix representations of the healthy subjects and the schizophrenia patients. Citing [33], the main difference between standard event related potentials and the method presented is the use of Stockwell features and the addition of a learning mechanism on top of the features. The results obtained for the model were: $88.7\% \pm 4\%$ discrimination accuracy rate, $100\%$ specificity rate, $77.4\% \pm 7.1\%$ sensitivity rate and $0.0078$ p-value significance.

There are several approaches to discriminate schizophrenia patients from healthy controls. Complexity and frequency analysis have been covered. Another way of looking at the EEG signal is through connectivity maps by performing an analysis of connectivity between nodes, as [34] did. The main differ-
ence between this method and conventional analysis is that the connectivity between nodes is measured in terms of the connectivity between each node and a third, reference node. So this is not a direct node analysis, [34]. In order to build these maps, a similarity function needs to be chosen, so one can check which nodes are more similar to which ones. The similarity function used in [34] was a correlation based one.

**Figure 3.3: Connectivity Measuring (sourced from [34]).**

The dataset used in [34] consists of 50 individuals, where 25 were diagnosed with schizophrenia and the other 25 were healthy. Each EEG recording has a total of 78 stimuli, presented for 150ms each with an inter-stimulus interval of 1s. Results showed that the degradation of connectivity is being accelerated within schizophrenia individuals. And that information relay changes in an abnormal manner primarily in the prefrontal area. In terms of accuracy, the connectivity maps method achieved 93.8\% ± 4\%; 100\% specificity rate; 87.6\% ± 5.1\% sensitivity rate; and 0.0041 significance p-value. This work gives a good insight on how connectivity maps can be applied to discriminate schizophrenia. And most important, that one should take into account that a change in a certain region can influence other regions in the brain.

In [35], another approach to classify schizophrenia was employed based on entropy and complexity measures of the EEG signal. The features extracted from the signal were: Shannon entropy, spectral entropy, approximate entropy, Lempel-Ziv complexity and Higuchi fractal dimension. Genetic programming was used for feature selection. With this features, two algorithms were implement to perform the classification: Adaptative boost (Adaboost) and Linear Discriminant Analysis (LDA). Both algorithms were validated with and without the feature selection step. The EEG dataset consisted of 20 schizophrenic and 20 healthy control individuals, with 20 channels. The population of individuals aged from 18 to 55. The recordings were done with eyes open. Without feature selection, the accuracy reached 0.8578 ± 0.1207
and $0.8961 \pm 0.1077$, for LDA and Adaboost respectively. With Genetic programming for feature selection, the accuracy reached $0.8875 \pm 0.0946$ and $0.9126 \pm 0.0924$. These results were obtained using leave-one-out cross validation.

The works introduced in this section yield extremely good results, although none of them use frequency features. In contrast, works in 3.1 provide very promising statistical analysis done on the frequency domain. This motivates this work, suggesting the need to further study the relevance of the frequency domain using state of the art algorithms.

3.3 Deep Learning

As mentioned before in chapter 2, deep learning is driving research towards new unprecedented possibilities. These models have the advantage of learning different properties of the data at different layers, but to do so one needs a sufficient amount of data in order for the network to generalize and actually learn. The lack of data can prevent long-awaited breakthroughs. In neuroscience, large datasets are hard to collect due to experimental setups and resources including subjects that satisfy the (typically strict) requirements. Although, techniques like EEG are way cheaper and easier to perform than Magnetic Resonance Imaging for example, it is still difficult to gather voluminous data to conduct specific tasks.

One way to tackle this problem is through Similarity Learning. A type of learning technique where the goal is to learn a model based on the observable distance or similarity between two instances. This kind of learning can be useful when the dataset is very small, and consequently when training on pairs of two or even three, the dataset size will increase as it is based on combinations. If the original dataset has $n$ samples, and your dataset of pairs has pairs of length $m$, then the dataset of pairs will have size $nC_m$. Similarity Learning not only is an alternative to have bigger data, but it also tackles the problem of few data to learn features as the learning objective is the distance/similarity between $m$ samples. The optimization function learns a distance space instead of a classification or regression one.

Siamese Neural Networks

First introduced by [36] as a novel model used in the task of signature classification, whose aim was to distinguish signature forgeries from the real ones, Siamese Neural Networks (SNN) are deep learning architecture that employ the same "siamese networks" for two inputs. The outputs of the pairs used as input to these "siamese networks" are joined in a distance function. In Fig. 3.4 one can see the architecture proposed. In [36], the proposed distance function between the output of the siamese networks, is the cosine similarity (for signatures from the same person the output should be 1, and –1 for forged ones). This model had outstanding results at the time, detecting $80.0\%$ of the forged signatures and $95.5\%$ of the genuine signatures.

More recently, [37] successfully used a Siamese Deep Architecture for One Shot Learning (meaning the model only sees each class once in an epoch). This approach had results, as good as, $92.8\%$ for
accuracy in the evaluation set. These results were achieved through a Siamese Convolutional Architecture, which, according to [37] is: 1) are capable of learning generic image features useful for making predictions about unknown class distributions even when very few examples from these new distributions are available; 2) easily trained using standard optimization techniques on pairs sampled from the source data; and 3) provide a competitive approach that does not rely upon domain-specific knowledge by instead extracting feature recurring to deep learning techniques. Once this kind of network is trained, its learned representations via a supervised metric-based approach with SNNs are useful to perform tasks like classification, relying on the discriminative properties of these features.

### 3.3.1 Deep Learning on Electroencephalography

As said before, Deep Learning techniques have emerged as state of the art and they are very well suited to detect patterns in data. That is why its use on EEG signal is so well received.

In [38], an attempt to use SNNs was made in the area of Music Information Retrieval, by ways of using the EEG signal recorded while users were listening to a certain audio sequence. Although the results were not statistically significant due to the difficult task at hand and the insufficient amount of data, the reason they applied SNNs is the same that motivates our work, which is the lack of data. This thesis is another attempt to use SNNs, but in a different area: Schizophrenia Classification; and with significantly more data, although not enough to train a standard deep learning model.

In [39], a Deep Recurrent-Convolutional architecture is proposed to process EEG signal based maps
of the brain. This work has the purpose to see how various connections made between the different topographic maps of different mental states vary. Citing [39], every individual has a different cognitive processing capacity which causally determines his/her ability in performing mental task. This work characterized mental load, approaching this problem from a classification stance.
Chapter 4

Solution

The problem being tackled in this work is: performing superior classification of schizophrenia through the analysis of EEG frequency domain. As mentioned before, the EEG signal properties make deep learning a high candidate to perform classification. And surveyed work on schizophrenia characterization and diagnosis shows the difficulty faced, due to noisiness and limited observations. On the other hand, the relevance of the frequency domain content attracts the use of deep learning techniques. To this end, a comprehensive statistical analysis is undertaken and a new class of neural network approaches for labeled electroencephalographic data is proposed. All of the analysis and tasks performed were done with the same dataset, [6]. This chapter further describes the considered methodology. In accordance, first a brief description of the dataset used in this work is presented, followed by four sections describing the proposed methodology: Frequency Analysis, Distance Analysis, Feature Extraction and Classification.

Figure 4.1: EEG channel setup.
Fig. 4.2 shows the pipeline of this chapter. Starting with Frequency Domain Analysis, following a Distance Analysis and the introduction to a Pairwise Distance Space. At last, another Distance Analysis approach is done followed by Advanced Diagnosis.

![Diagram of the pipeline](image)

**Figure 4.2: Approach pipeline.**

### 4.1 Preprocessed Data

The dataset has a total of 84 individuals: 39 healthy controls and 45 schizophrenia diagnosed subjects. This population consists of adolescents who had been screened by psychiatrists and diagnosed or not with the schizophrenia neuropathology. Figure 4.1 has the channel setup used for the recording sessions. The original description of the dataset can be found in [6].

### 4.2 Frequency Analysis

To perform an analysis in the frequency domain, first the discrete fourier transform on each of the 16 channels of each of the 84 individuals is taken. This discrete fourier transform is done using the fast fourier transform (FFT) algorithm. An illustrative result of a fast fourier transform is shown in Fig. 4.3. Further, one can use the DSTFT to compute an image, as shown in Fig. 4.4.

![FFT algorithm over a 1 minute EEG channel recording](image)

**Figure 4.3: FFT algorithm over a 1 minute EEG channel recording.**

This representation has a very thin granularity with over 1900 frequency positions (every 0.1Hz).
EEG frequency analysis can be done in multiple ways, specifically when tuning the frequency granularity, in other words the interval of frequencies one wants to take into account. According to related work, frequency analysis is usually done over an interval of values, taking the sum of the magnitude of each frequency in a certain interval, also known as frequency bands as introduced in Chapter 2. This makes the computation faster and although the resolution is worst, the interval sum gives a good generalization to compare individuals. This reasoning is described in equation 4.1, where $b$ is a frequency band according to a certain frequency division set $B$, $l$ and $u$ are the lower and upper bounds of the interval of the band, and $I$ is the magnitude given by the FFT.

$$b \in B : b = \sum_{i} I_i$$

(4.1)

The frequency bands used are: delta ($0.5 < \delta \leq 4$ Hz), theta1 ($4 < \delta \leq 6$ Hz), theta2 ($6 < \delta \leq 8$ Hz), alpha1 ($8 < \delta \leq 10$ Hz), alpha2 ($10 < \delta \leq 13$ Hz), beta1 ($13 < \delta \leq 16$ Hz), beta2 ($16 < \delta \leq 23$ Hz), beta3 ($23 < \delta \leq 30$ Hz) and gamma ($30\text{Hz} < \delta \leq 128$ Hz). These intervals were applied taking into account the analysis done in [8], which claims that this division is relevant for schizophrenia diagnosis.

The frequency analysis is going to be done: globally (taking into account all channels and frequency bands); by channel; and by frequency band. In the global analysis, first the frequency bands are summed and then, for each individual, the sum of the frequency bands is averaged over the channels producing $f$, where $f = \frac{1}{N} \sum_{i=1}^{M} \left( \sum_{i}^{F} I_i \right)$, being $N$ the number of channels, $F$ the number of frequencies in each channel and $I$ the magnitude of each frequency. Channel-specific (by channel) analysis returns as many as many distributions as the number of channels, since the frequencies are averaged over the whole population for each channel (i.e. a single power spectrum per channel and population). In the frequency band analysis, each band in each channel is averaged over the whole population.

### 4.3 Distance Analysis

Given the rich spectral content of the EEG signal, it is important to understand how it differs between different populations (healthy controls and schizophrenian individuals). A distance and similarity analysis is going to be performed, the goal is to perform an analysis where to understand the baseline spectral content and see how the data is distributed. Euclidean distance, cosine similarity and dynamic time warping are the appleid distance metrics. These metrics are going to be used with the following features: the output of the fast fourier transform; output of the discrete short-time fourier transform; and different versions of the last two with granularity differences (e.g., instead of using the original frequency granularity, frequency band intervals are used).

The euclidean distance is used as a baseline distance to compare between the two other ones. The cosine similarity is used with the motivation that schizophrenia individuals and healthy controls can be better differentiated in terms of correlation (angle/similarity between each other) than with distance magnitude differences, which captures the differences between different frequency bands without the effect of magnitude of each frequency. Dynamic Time Warping gives the time variation of frequency
waves importance to see how individuals vary.

For each distance metric a distance matrix is built (with the rows and columns being the individuals and the values of each cell the distance between two individuals). Once this matrix is computed, several plots are obtained to see how the data behaves.

4.3.1 Distances applied on DFT: Euclidean Distance and Cosine Similarity

This distances are taken in the following setups: globally; by channel; by frequency band; by channel and frequency band.

Global

The metric is performed over the FFT output, which has the magnitudes, $F$, of the frequencies. With this value, for two individuals, the distance/similarity between each channel is averaged over the channels. This reasoning is shown in equation 4.2, where $X$ and $Y$ are two individuals, $X_i$ and $Y_i$ is the FFT of channel $i$ of each individual out of a total of $M$ channels, finally $D_f$ is the distance function. This value is placed in the distance matrix at the respective cell of the two individuals.

$$D(X,Y) = \frac{1}{M} \sum_{i=1}^{N} D_f(X_i, Y_i)$$  \hspace{1cm} (4.2)

By channel

In this setup, there are as many distance matrices as the number of channels. The goal is to check if there are channels with more relevant differences than others (important step for feature selection). For each channel, the distance/similarity between the FFT output is performed and placed in the respective distance matrix of that channel. The difference between this and the global approach, is that in by channel the distances of each channel are not averaged, as shown in equation 4.3.

$$i \in [1, M] : D_i(X,Y) = D_f(X_i, Y_i)$$  \hspace{1cm} (4.3)

By frequency band

Using the original granularity of the FFT, this setup computes the distance/similarity in the interval of the frequency band. There are as many distance matrices as frequency bands. After the distance/similarity for all channels is computed, each frequency band is averaged over the channels, as shown in equation 4.4, where a band $b$ as a lower and upper bound values of frequency, delimiting the interval.

$$b \in B : D_b(X,Y) = \frac{1}{M} \sum_{i=1}^{M} D_f(X_i^b, Y_i^b)$$  \hspace{1cm} (4.4)
By channel and frequency band

For each channel and frequency band, the distance/similarity is computed just as it is computed in the by frequency band setup, but the average is not taken, as shown in equation 4.5. With this procedure one will have $M \times \#B$ distance matrices.

$$b \in B \ \& \ i \in [1, M] : D^b_i(X, Y) = D_f(x^b_i, y^b_i)$$  \hspace{1cm} (4.5)

4.3.2 Distances applied on DSTFT: Dynamic Time Warping

The motivation behind the use of the Dynamic Time Warping algorithm, is to see if the frequency bands variation through time have a significance that differentiates between healthy controls between schizophrenian individuals.

![Figure 4.4: Discrete Short-time Fourier Transform of a 1 minute EEG channel recording.](image)

The Discrete Short-time Fourier Transform is applied to the 1 minute EEG of each channel, and the magnitude of the frequencies between each frequency band interval are summed. In the end, the time resolution is the same as the output of the DSTFT (illustrated in Fig. 4.4), but the frequency resolution will be the same as the number of frequency bands taken into account.

The analysis of the distances provided by the DTW algorithm can be made in terms of channel, frequency band and both. Further the analysis will be more concentrated channel wise, as a means of feature selection procedure for a deep learning model.

4.4 Siamese Neural Network

The Siamese Neural Network (SNN) is built so one can extract distance learned features out of the EEG signal, and with this features apply a classification algorithm to diagnose schizophrenia. The model is inspired by the Convolutional Neural Network version in [37] and its inputs are pairs of the outputs of each channel of the DSTFT. Although training one network for each EEG channel is the best way to develop the model, because it is not mixing different channels that can be seen as different spaces.
Unfortunately, due to the lack of data, the results obtained in this setting were very poor. Taking into account two bullet points: 1) Data augmentation can be performed in one dataset by the addition of noise to each sample; and 2) each EEG channel is highly correlated between channels that are close to it, with only slight differences in terms of frequency. With this, the mixture of all of the 16 channels in the same dataset to train the model was made, as a need to get more data and justified by a type of data augmentation technique (adding noise to the dataset), by no means of any data augmentation algorithm but by mixing channels. On top of this, as mentioned in 3.3, this type of networks are capable of learning generic image features useful for making predictions about unknown class distributions even when very few examples from these new distributions are available. So one is capable of saying that this networks can learn the different spaces/distributions of each EEG channel. In [23], a Siamese Architecture was applied and achieved great results in faces never before seen, which encourages even more the use of this type of architecture in this work.

Summarizing the above paragraph, the proposed approach offers this six major contributions:

1. Ability to learn from small datasets by taking advantage of Siamese Network layering, inherently prepared to extract features from a limited number of observations of a possible data distribution under study [6]. The features produced by these networks have proven to be useful to perform classification as they rely on either the homologous or discriminative properties of observation-pairs in a pairwise distance domain [37];

2. Ability to deal with the rich and complex spectral and temporal content of EEG data by processing the signal into spectral images with a fine frequency and temporal resolution per electrode, and by subsequently reshaping the Siamese network architecture with adequate convolutional operations;

3. Robustness to noise and wave-instability by assessing distances on the spectral content under a cosine-loss. Gathered evidence show less susceptibility to artefacts and the inherent variability of electrophysiological potentials composed of continuously changing overlapping electrical fields created by localized neurons [5];

4. Ability to deal with the multivariate nature of the signal (rich spatial content) by capturing inter-dependencies between channels as their content is simultaneously used to shape the weights of shared connections in the network;

5. Ability to handle the extremely-high dimensional nature of the gathered spectral content of brain signal data (high-resolution spectral image per scalp location) under L1 regularization;

6. Applicability of the proposed EEG-based diagnostics to alternative populations or diseases, evidenced by the: i) placed Bayesian optimization step [26] for hyperparameter tuning and fixing feature numerosity; ii) fully-automated nature of the approach once signals are recorded; and iii) generalization ability of the learning process on validation data.
4.4.1 Dataset Structure

So to understand how our dataset is structured, if it consisted of only two individuals with 16 EEG channels each, one will have 16 pairs corresponding to each channel to train the neural network, being a pair \( i \in [1, N] : P_i = (X_i, Y_i) \), defined by \( P \). As our dataset has 84 individuals, the size will be higher by combining every pair of individuals possible, though only equal channels are paired. The input of the model is intended to be a pair of the DSTFT representation of the EEG signals, just as illustrated in Figure 4.4. The DSTFT is taken with windows of 2 seconds so one can reach frequencies as low as 0.5 Hz.

![Figure 4.5: Architecture of the base neural network.](image)

![Figure 4.6: Architecture of the neural network used to train the weights for the base network.](image)
4.4.2 Network Architecture

The architecture of the base neural network consists of two Convolutional layers, both using ReLU as activation function. After each Convolutional Layer a Dropout of 50% is introduced, this value is the one recommended to use by [20]. All this followed by a fully-connected softmax activation layer, which outputs a feature vector representation of the DSTFT. This layout of the architecture proposed can be seen in Figure 4.5. Further, $L_1$ regularization is added to the kernel of each layer and is useful because it helps remove features that are not useful for the task. Adam [25] is used to optimize the network during the training session.

The base network is replicated and a twin of it, or in other words, its "siamese" will join it (with their outputs being connected by a distance function) and accomplish the SNN, whose final output will be the value of the distance between the outputs of the base networks. This architecture layout can be seen in Figure 4.6. The SNN will learn to separate instances of schizophrenic and healthy individuals, by means of the cosine distance and the contrastive loss.

4.4.3 Hyperparameter Tuning

The number of layers, as well as their type, are fixed. The rest of hyperparameters (regularization factor, margin value, learning rate, kernel size and output dimension of the BN) are susceptible to optimization. BO is set to run with a maximum of 250 acquisitions and starts with 5 iterations to perform an initial exploration. In each iteration and acquisition, a $K$-fold Cross Validation with $K = 5$ is done with the whole dataset. The combination of hyperparameters that has the best average validation accuracy across the 5-folds is chosen to perform the feature extraction. Each of the hyperparameters are assigned the following value domains to explore: regularization factor $\in [10^{-3}, 10^{-1}]$, margin value $\in [1.0, 2.0]$, learning rate $\in [10^{-6}, 10^{-3}]$, kernel size $t \times f$ with $t = f = \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ (the same kernel size is used for both convolutional layers) and final output dimension $\in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14\}$. The BO surrogate model is a standard Gaussian Process. Expected Improvement is used as an acquisition function and the Limited-Memory Broyden–Fletcher–Goldfarb–Shanno algorithm as the acquisition optimizer.

The DSTFT magnitudes are normalized, under the hypothesis that there exists a threshold from which there is no additional information to identify the schizophrenia pathology. With this, the values are normalized by an upper value, $U$. Values of $f$ smaller than $U$ are divided by $U$ and magnitudes bigger than $U$ are set to 1.0. This allows every magnitude of the frequencies to be within the interval $[0, 1]$ after the normalization is performed. We take advantage of the BO exploration to obtain $U$, by introducing it in the same optimization process made for the SNN hyperparameters. The domain assigned to be explored for $U$ is $[100.0, 500.0]$.

The transformation performed by the base network can be seen as in equation 4.6, where the $W$ and $B$ are the only set of learnable parameters, being $X$ the DSTFT input of a channel of an individual and $Y$ the features given by this transformation.

$$Y = WX + B$$ (4.6)
In the end, the base network will be extracted after the training based on pairs is done, and will serve as a feature extraction transformation technique to obtain the distance based features that should be able to perform a good separation between schizophrenian individuals from healthy controls.

4.5 Hierarchical Clustering and Community Analysis

To compare how the different sets of features (FFT, DSTFT and the features learned from the previously encountered pairwise distance space) behave, we compare their performance by means of analysis through hierarchical clustering techniques (dendogram algorithm) and community finding algorithms. Specifically, we consider Girvan Newman’s as it gives different partitions of a network and can be seen as hierarchical partitions along its different iterations. With these approaches, a visualization of the data is given and, depending on its consistency, it proves how good each distance and each set of features is optimal to use in a schizophrenia classification task. The goal is to obtain a clear separation from healthy controls and schizophrenian individuals, when taking into account the whole dataset.

These algorithms are also useful to check the different types of schizophrenian individuals that are present in the dataset. Under the application of this clustering analysis on the subset of schizophrenia individuals, one should be able to assess whether it is possible to identify subtypes of schizophrenia diagnosed, which correspond to the types of schizophrenia diagnosed in the psychiatry sessions. These types are namely: infant schizophrenia, schizotypical and schizoaffective disorders.

4.6 Classification

After the SNN has been trained (in a 20 epochs session), the outputs of the BN for every example were the result of our feature extraction process. With these features, the following classifiers were trained to identify schizophrenia: Support Vector Machines (SVM), Random Forest (RF), XGBoost (XGB), Naive Bayes (NB) and k-Nearest Neighbors (kNN). This process was performed with a Leave-One-Out Cross Validation (LOOCV). For each of these classifiers, BO hyperparameter tuning is also performed, setup with a maximum of 10 acquisitions and 5 iterations for initial exploration. The hyperparameter domains for each classifier were:

- SVM: type of kernel (linear or radial-basis function kernel), cost $C \in [0.5, 5]$, and gamma coefficient $\gamma \in [0.00001, 1.0]$
- RF: number of estimators $N_e \in \{5, 10, 15, 20, 25\}$
- XGB: maximum depth $d \in \{3, 4, 5, 6, 7\}$, learning rate $\lambda \in [0.001, 0.1]$, and number of estimators $N_e \in \{10, 50, 100, 200\}$
- NB has no hyperparameters
- kNN: number of neighbors $k \in \{2, 3, 4, 5, 6, 7, 8\}$
The hyperparameter tuning optimization for the classifiers is also performed in a $K$-Fold Cross Validation setup ($K = 5$), but instead of using the whole dataset (as was the case for the SNN) only the training set of the LOOCV partition was used. Similar to the BO for the SNN, the combination of hyperparameters with the best average validation accuracy is chosen for each classifier.
Chapter 5

Results

In this chapter, the results obtained using the proposed methodology in Chapter 4 are presented. In accordance, Section 5.1 shows what was obtained with the Frequency Analysis and Distance Analysis referred in Chapter 4. Section 5.2 has the results for the Classification task and Hierarchical Clustering using Frequency Features. Section 5.3 has the explanation of the siamese model structure. Section 5.4 has the results for the Classification task and Hierarchical Clustering using the extracted features from the siamese model. At last, Section 5.5 has an interpretation from all the contents presented in this chapter.

5.1 Data Analysis

5.1.1 Frequency Analysis

![Delta wave activity distribution in channel T3](image)

Figure 5.1: Example of how the distributions were analyzed.

To perform a frequency analysis over the dataset, two sets of individuals are defined:
• \(hc\) - The set of healthy control individuals.

• \(scz\) - The set of schizophrenian individuals.

Fig. 5.1 shows how each distribution was treated and obtained. The horizontal axis refers to the frequency band intensity and the vertical axis provides the proportion of the population with that amount of intensity. Normal refers to the absence of transformations in scale, as some transformations were made (Log-scale transformations, to check how the data behaved). The granularity chosen was the frequency bands defined in Chapter 4.

It was found that the magnitudes of the frequencies varied between each channel and each frequency band. Further, the differences between healthy controls and schizophrenian individuals were explored via channel and band. The results obtained are shown in Fig. 5.2.

5.1.2 Distance Analysis

In this section, four sets are defined:

• \(hc\) - The set of distances between healthy control individuals, \(hc = \{d_{ij} | i = 1..39, j = 1..39, d_{ij} = d(x_i, x_j)\}\).
- \textit{scz} - The set of distances between schizophrenian individuals, \( \text{scz} = \{ d_{ij} | i = 40..84, j = 40..84, d_{ij} = d(x_i, x_j) \} \).

- \textit{hc} and \textit{scz} - The set of distances between schizophrenian individuals plus the set of distances between healthy control individuals - represented by \( \text{hc} \cup \text{scz} \).

- \textit{hc} vs \textit{scz} - The set of distances between schizophrenian and healthy control individuals.

The purpose of the \textit{hc} and \textit{scz} sets is to check how much variance is within healthy controls and schizophrenian individuals. The \textit{hc} and \textit{scz} sets are defined to see how different distance metrics perform in those sets. If the values between the last two sets are different, it means they can be separated with a certain distance measure.

The Figures that follow this paragraph consist of the distribution of the distances according to each subset of the population pairs set.

**Euclidean Distance**

![Figure 5.3: Euclidean distances between individuals of each set using FFT original granularity.](image)

![Figure 5.4: Euclidean distances between individuals of each set using frequency band granularity.](image)

(a) Normal distances scale.

(b) Log-scaled distances.
Figure 5.5: Euclidean distances on $he_{and\_scz}$ and $he_{vs\_scz}$ sets.

Cosine Distance

Figure 5.6: Cosine distances between individuals of each set using FFT original granularity.

Figure 5.7: Cosine distances between individuals of each set using frequency band granularity.
In Figs. 5.5 and 5.8 it is shown the difference by percentage of $hc_{and}_{scz}$ and $hc_{vs}_{scz}$ vary by channel, using FFT original granularity and frequency band granularity, but still it is a low percentage value. As seen in Figs. 5.3, 5.4, 5.6 and 5.7 the defined sets are not well differentiated by neither the Euclidean Distance nor the Cosine Distance. DTW on the DSTFT of the signals was further applied in accordance with the principles introduced in Chapter 4.

**Dynamic Time Warping**

A next approach was performed using the DTW algorithm.

Taking Fig. 5.9 observations into account, the DTW distance approach performed worse against the other approaches.

With this section, one can be able to say that the FFT frequency features have a low performance when one uses them to differentiate healthy controls from schizophrenian individuals.
5.2 Frequency Features

This section dedicates itself to analyze the relevance of FFT frequency features to guide the clustering and classification of individuals. A hierarchical clustering analysis is first shown followed by the performance of classifiers.

5.2.1 Hierarchical Clustering

Dendrogram

From Figs. 5.10 - 5.18, the Cosine Distance is the metric that shows the best performance when using the Dendrogram algorithm, if one analyzes Fig. 5.13 it can be seen that a clear cut can be made with a perfect separation of healthy controls from schizophrenic individuals.

Community Finding - Girvan Newman Algorithm

Figure 5.19: Network Analysis, edges weighted with Euclidean Distance.

Figure 5.20: Network Analysis, edges weighted with Cosine Distance.
Figure 5.10: Dendrogram with all of the individuals (Euclidean Distance).
Figure 5.11: Dendrogram with all of the healthy control individuals (Euclidean Distance).
Figure 5.12: Dendrogram with all of the schizophrenian individuals (Euclidean Distance).
Figure 5.13: Dendrogram with all of the individuals (Cosine Distance).
Figure 5.14: Dendrogram with all of the healthy control individuals (Cosine Distance).
Figure 5.15: Dendrogram with all of the schizophrenian individuals (Cosine Distance).
Figure 5.16: Dendrogram with all of the individuals (DTW Distance).
Figure 5.17: Dendrogram with all of the healthy control individuals (DTW Distance).
Figure 5.18: Dendrogram with all of the schizophrenian individuals (DTW Distance).
From the three Figs. 5.19, 5.20 and 5.21, the Cosine Distance is the metric that shows the best layout done by the Fruchterman-Reingold force-directed algorithm. But the GN algorithm cannot get good partitions with equally a balanced number of healthy controls and schizophrenic individuals, producing communities with single individuals and one big community with the rest of the population. This happens for the Cosine, Euclidean and DTW distances for the FFT and DSTFT features to compare later with the feature learned space.

### 5.2.2 Classification

The results of the classification task for the FFT features with the original granularity are shown in Table 5.1, with *p-value*, $p$.

Table 5.1: Leave One Out Cross Validation on FFT frequency extracted features with FFT granularity.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT-SVM</td>
<td>0.66 ± 0.28</td>
<td>0.69 ± 0.26</td>
<td>0.63 ± 0.29</td>
</tr>
<tr>
<td>FFT-NB</td>
<td>0.57 ± 0.32</td>
<td>0.33 ± 0.38</td>
<td>0.85 ± 0.14</td>
</tr>
<tr>
<td>FFT-kNN</td>
<td>0.60 ± 0.31</td>
<td>0.56 ± 0.33</td>
<td>0.64 ± 0.30</td>
</tr>
<tr>
<td>FFT-RF</td>
<td>0.58 ± 0.32</td>
<td>0.58 ± 0.32</td>
<td>0.64 ± 0.29</td>
</tr>
<tr>
<td>FFT-XGB</td>
<td>0.65 ± 0.28</td>
<td>0.68 ± 0.26</td>
<td>0.61 ± 0.30</td>
</tr>
</tbody>
</table>

Further, the FFT features with the frequency band granularity previously introduced are shown in Table 5.2.

Table 5.2: Leave One Out Cross Validation on FFT frequency extracted features with band granularity.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFB-SVM</td>
<td>0.54 ± 0.34</td>
<td>1.00 ± 0.00</td>
<td>0.00 ± 0.00</td>
</tr>
<tr>
<td>FFB-NB</td>
<td>0.57 ± 0.32</td>
<td>0.42 ± 0.38</td>
<td>0.74 ± 0.22</td>
</tr>
<tr>
<td>FFB-kNN</td>
<td>0.52 ± 0.34</td>
<td>0.55 ± 0.33</td>
<td>0.40 ± 0.36</td>
</tr>
<tr>
<td><strong>FFTB-RF</strong></td>
<td><strong>0.65 ± 0.28</strong></td>
<td><strong>0.67 ± 0.27</strong></td>
<td>0.63 ± 0.30</td>
</tr>
<tr>
<td>FFB-XGB</td>
<td>0.63 ± 0.29</td>
<td>0.64 ± 0.29</td>
<td>0.61 ± 0.30</td>
</tr>
</tbody>
</table>
5.3 Siamese Model

In this section, the hyperparameters obtained in a Bayesian Optimization (BO) session are shown. Following, a Convolutional Neural Network (CNN) with the same architecture as the Base Network (BN), the base network of the Siamese Neural Network (SNN), is introduced in order to justify the use of the SNN versus a standard CNN classifier.

5.3.1 Hyperparameter Tuning

Once the BO hyperparameter optimization process finished the values for the hyperparameters of the SNN were: $1.49586563 \times 10^{-3}$ for the learning rate; a window of size $5 \times 11$ for the kernel; $10$ for the final vector dimension (output of the network); $1.65531895 \times 10^{-2}$ as the regularization constant $\lambda$; $1.18389688$ as the margin value for the contrastive loss; and $104.842448$ as the normalization factor for the DSTFT features. These suboptimal hyperparameters got a 5-Fold Cross Validation average accuracy of $0.71$.

The same process was done for the CNN classifier, using BO with $250$ acquisitions and $5$ initial exploration iterations. The hyperparameters of the SNN were: $0.0014295145526949326$ for the learning rate; a window of size $11 \times 4$ for the kernel; $8$ for the final vector dimension (output of the network); $0.011611445489506106$ as the regularization constant $\lambda$; and $295.9626437499775$ as the normalization factor for the DSTFT features. These suboptimal hyperparameters got a 5-Fold Cross Validation average accuracy of $0.85$. The CNN was not subject to optimize the margin factor as it is not integrated in the model. The rest of the hyperparameters had the same domains used for exploration in the BO for the SNN.

5.4 Similarity Learned Features

Once the SNN was subject to BO process to obtain the respective hyperparameters, an analysis of how the extracted features from the Base Network of the SNN is done. This analysis is done in two main ways, just as it was done for the standard cosine and euclidean distance for the FFT features: Hierarchical Clustering and Classification.

5.4.1 Hierarchical Clustering

To check how the features, obtained by the siamese model learned transformation perform on an Hierarchical Clustering setting, the dendogram and GN algorithms are applied to the distances between all of the individuals (training + testing set).

Dendogram

Figs 5.22, 5.23 and 5.24 refer to the dendogram representations of the respective populations using the Cosine Distance with the SNN extracted features.
Figure 5.22: Dendrogram with all of the individuals (SNN features with Cosine Distance).
Figure 5.23: Dendogram with all of the healthy control individuals (SNN features with Cosine Distance).
Figure 5.24: Dendrogram with all of the schizophrenic individuals (SNN features with Cosine Distance).
Community Finding - Girvan Newman Algorithm

(a) Network representation  (b) Network communities (each community is assigned a color and the number of each edges represents the individual id).

Figure 5.25: Network Analysis, edges weighted with Cosine Distance on the SNN extracted features.

5.4.2 Classification

Table 5.3: Leave One Out Cross Validation on the Similarity learned features.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT-SVM</td>
<td>0.66 ± 0.28</td>
<td>0.69 ± 0.26</td>
<td>0.63 ± 0.29</td>
</tr>
<tr>
<td>FFT-NB</td>
<td>0.57 ± 0.32</td>
<td>0.33 ± 0.38</td>
<td>0.85 ± 0.14</td>
</tr>
<tr>
<td>FFT-kNN</td>
<td>0.60 ± 0.31</td>
<td>0.56 ± 0.33</td>
<td>0.64 ± 0.30</td>
</tr>
<tr>
<td>FFT-RF</td>
<td>0.58 ± 0.32</td>
<td>0.58 ± 0.32</td>
<td>0.64 ± 0.29</td>
</tr>
<tr>
<td>FFT-XGB</td>
<td>0.65 ± 0.28</td>
<td>0.68 ± 0.26</td>
<td>0.61 ± 0.30</td>
</tr>
<tr>
<td>DSTFT-CNN</td>
<td>0.76 ± 0.21</td>
<td>0.76 ± 0.21</td>
<td>0.77 ± 0.20</td>
</tr>
<tr>
<td>DSTFT-SVM</td>
<td>0.82 ± 0.16</td>
<td>0.84 ± 0.14</td>
<td>0.79 ± 0.18</td>
</tr>
<tr>
<td>DSTFT-NB</td>
<td>0.83 ± 0.16</td>
<td>0.84 ± 0.14</td>
<td>0.81 ± 0.17</td>
</tr>
<tr>
<td>DSTFT-kNN</td>
<td>0.82 ± 0.16</td>
<td>0.84 ± 0.14</td>
<td>0.79 ± 0.18</td>
</tr>
<tr>
<td><strong>DSTFT-RF</strong></td>
<td><strong>0.84 ± 0.14</strong></td>
<td><strong>0.87 ± 0.12</strong></td>
<td><strong>0.82 ± 0.17</strong></td>
</tr>
<tr>
<td>DSTFT-XGB</td>
<td>0.82 ± 0.16</td>
<td>0.84 ± 0.14</td>
<td>0.79 ± 0.18</td>
</tr>
</tbody>
</table>

In Table 5.3, the results obtained with the classifiers considered are compared between each other and with the FFT features as well, and also compared with the CNN classifier.

5.5 Discussion

Section 5.1 results indicate that the FFT frequency based features do not have a clear discriminative power over Schizophrenia. Regarding the ability to distinguish the populations, Section 5.1.2 results suggest that DTW and Euclidean distances are limited in this context. Yet, in contrast, the Cosine distance is able to get what one could say a good separation of the data when applying the dendogram algorithm. Facing this behaviour with the Dendogram, one expected to see good results with the classifiers. The latter was not true. According to Tables 5.1, 5.2 and 5.3, the SNN features outperform the FFT frequency baselines considered by an average of 19pp in accuracy. One particular observation is that
among the FFT features there was a clear difference between classifiers: FFT-kNN and FFT-NB were worse than FFT-XGB, FFT-RF and FFT-SVM. The same was not found among the SNN features, with all of the variants performing comparably. This is due to most of the discriminative work of the problem being solved by the SNN transformation of the DSTFT representation of the EEG signal. These observations suggest that the SNN alone is capable of generalizing better and have extremely discriminative features.

Further, as expected the DSTFT-CNN baseline underperformed by an average of $-8\text{pp}$ than all of the classifiers that used SNN features. The latter is justified by the low amount of data provided to train the model, which in contrast is much smaller than the amount of data used to train the SNN model. Adding to this, the high dimensionality of the DSTFT features is a characteristic that DSTFT-CNN is not originally prepared to handle. On the other hand, the regularization techniques still present in the DSTFT-CNN baseline were key factors to outperform the FFT baselines. With the DSTFT-CNN baseline, the SNN-RF got a p-value of $0.000000003$ for accuracy, $0.00006$ for sensitivity and $0.0004$ for specificity. The DSTFT-CNN was chosen so one could compare different deep learning approaches.

There was an expectation about the bad results obtained in the DSTFT-CNN model being due to a vanishing gradient, but tests were done and Batch Normalization was introduced to tackle this problem, and the same results were obtained.

From the Hierarchical Clustering techniques applied, the analysis was shown to be able to distinguish healthy from schizophrenic individuals, while in contrast the GN algorithm was unable to recover these two communities. This might be due to the way the edges were chosen to make the partition. Overall, the Dendograms under a Cosine distance with either the original FFT features or the extracted SNN features hold better results to discriminate the Schizophrenia pathology. As a final remark, we expected the DTW algorithm to adequately handle the inherent spectral variability between individuals, yet results show that it is not competitive against the Cosine distance.
Chapter 6

Conclusions

This thesis presented an analysis and different techniques to handle EEG signal, in order to perform a Schizophrenia diagnostic. Every single one of this techniques to handle the EEG signal, worked on the frequency domain of it, either through the FFT or DSTFT tranformation from the time domain. These two representations were subject to different approaches. Promising results were found using the Dendogram algorithm to perform an Hierarchical Clustering analysis using the Cosine Distance. Unfortunately, the latter is not a scalable solution when presented with high amounts of data. Notable examples of connectionist and spectral approaches have been proposed to discriminate and characterize Schizophrenia. Nevertheless, there is still a research gap on how to simultaneously explore the rich spectral, temporal and spatial nature of brain signals to perform classification. In spite of the indisputable role of neural network learning for the analysis of complex spatiotemporal signal data, its role for EEG-based diagnostics of psychiatric disorders remains largely unexplored due to the absence of large cohorts and the inherent stochastic complexities associated with electrophysiological data.

Once an initial analysis of the data was done, a motivation had risen to apply deep learning techniques to the data available. The rich nature of the electrophysiological data measured at the cerebral cortex makes deep learning a natural candidate to study disorders disrupting the normal brain activity. Nevertheless, the limited size of case-control populations, together with the inherent variability of the spectral content within and among individuals, had left the value of neural network approaches largely unexplored. By reshaping the architecture, loss and applied regularization, we show that the use of neural networks to classify Schizophrenia can increase the accuracy of diagnostics by 15-to-20 percentage points against peer alternatives (without hampering sensitivity or specificity). Two master principles underlie these results: 1) the mapping of the original data space into a pairwise distance space to support data augmentation while enhancing the discriminative power of the output features; and 2) the exploration of the rich nature of brain patterning through convolution operations on the spectral imaging of the signal, with weights learned under a cosine loss to better account for the inherent noise of electrophysiologic data.
6.1 Future Work

As future work:

- we aim to extend the experimental analysis towards alternative disorders and populations with potentially different EEG instrumentation or protocol;

- contrast the performance of the proposed EEG-based learners against state-of-the-art brain imaging learners on a population of individuals with (and without) neurodegenerative conditions being currently monitored;

- establish a method that is capable of performing a neurofeedback technique to tackle Schizophrenia symptoms, similarly to what has been previously proposed [1].
Bibliography


