Università di Genova

DIPARTIMENTO DI
INFORMATICA, BIOINGEGNERIA,
ROBOTICA E INGEGNERIA DEI SISTEMI

Machine Learning for Understanding Focal Epilepsy

Vanessa D'Amario

Theses Series

DIBRIS-TH-2020-17

Università di **Genova**

Dipartimento di Informatica, Bioingegneria,
Robotica ed Ingegneria dei Sistemi

Ph.D. Thesis in
Computer Science and Systems Engineering
Computer Science Curriculum

Machine Learning for Understanding Focal Epilepsy

by

Vanessa D'Amario

May, 2020

Ph.D. Thesis in Computer Science and Systems Engineering (S.S.D. INF/01)
Dipartimento di Informatica, Bioingegneria,
Robotica ed Ingegneria dei Sistemi
Università di Genova

Candidate
Vanessa D'Amario
vanessa.damario@dibris.unige.it

Title
Machine Learning for Understanding Focal Epilepsy

Advisors
Annalisa Barla
DIBRIS, Università di Genova
annalisa.barla@unige.it

Alessandro Verri
DIBRIS, Università di Genova
alessandro.verri@unige.it

External Reviewers
Thomas Moreau
Parietal INRIA Saclay, Ecole Polytechnique, Paris
thomas.moreau@inria.fr

Annalisa Pascarella
Istituto per le Applicazioni del Calcolo (IAC) "M. Picone"
Consiglio Nazionale delle Ricerche (CNR - National Research Council)
a.pascarella@iac.cnr.it

Location
DIBRIS, Univ. di Genova
Via Opera Pia, 13
I-16145 Genova, Italy

Submitted On
May 2020

A mia madre
Io sono qui.

Abstract

The study of neural dysfunctions requires strong prior knowledge on brain physiology combined with expertise on data analysis, signal processing, and machine learning.

One of the unsolved issues regarding epilepsy consists in the localization of pathological brain areas causing seizures. Nowadays the analysis of neural activity conducted with this goal still relies on visual inspection by clinicians and is therefore subjected to human error, possibly leading to negative surgical outcome.

In absence of any evidence from standard clinical tests, medical experts resort to invasive electrophysiological recordings, such as stereoelectroencephalography to assess the pathological areas. This data is high dimensional, it could suffer from spatial and temporal correlation, as well as be affected by high variability across the population. These aspects make the automatization attempt extremely challenging.

In this context, this thesis tackles the problem of characterizing drug resistant focal epilepsy. This work proposes methods to analyze the intracranial electrophysiological recordings during the interictal state, leveraging on the presurgical assessment of the pathological areas.

The first contribution of the thesis consists in the design of a support tool for the identification of epileptic zones. This method relies on the multi-decomposition of the signal and similarity metrics. We built personalized models which share common usage of features across patients.

The second main contribution aims at understanding if there are particular frequency bands related to the epileptic areas and if it is worthy to focus on shorter periods of time. Here we leverage on the post-surgical outcome deriving from the Engel classification. The last contribution focuses on the characterization of short patterns of activity at specific frequencies.

We argue that this effort could be helpful in the clinical routine and at the same time provides useful insight for the understanding of focal epilepsy.

Publications

'Hey there's DALILA: a DictionAry LearnIng LibrAry', Veronica Tozzo, Vanessa D'Amario, Annalisa Barla, *2017 Imperial College Computing Student Workshop (ICCSW 2017)*, (2018), Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

'Classification of Epileptic Activity Through Temporal and Spatial Characterization of Intracranial Recordings', Vanessa D'Amario, Gabriele Arnulfo, Lino Nobili, Annalisa Barla. *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, (2018) pp. 69–79, Springer.

'Multi-task multiple kernel learning reveals relevant frequency bands for critical areas localization in focal epilepsy', Vanessa D'Amario, Federico Tomasi, Veronica Tozzo, Gabriele Arnulfo, Lino Nobili, Annalisa Barla. *Machine Learning for Healthcare Conference* (2018), pp. 348–382.

In preparation for BMC bioinformatics **'Data Driven Interpretable Feature Selection for the Characterization of Epileptogenic Zones'**, Vanessa D'Amario, Gabriele Arnulfo, Lino Nobili, Giorgio Lo Russo, Alessandro Verri, Annalisa Barla. *Machine Learning for Healthcare Conference* (2018), pp. 348–382.

Acknowledgements

I want to thank my supervisors Professor Annalisa Barla and Professor Alessandro Verri for the support given throughout all these years.

A special thanks to Veronica Tozzo for having been a great friend, supporter, collaborator, and also a great informal reviewer of this thesis. Still thanks to Veronica and Federico Tomasi, for having been amazing coauthors, working together was fun.

Thanks to Gabriele Arnulfo, PhD, and Lino Nobili, MD, for their professional assistance and their precious suggestions.

I am grateful to Ospedale Ca' Granda Niguarda, and the Center of Sleep, for having me provided with the dataset analyzed through this thesis.

Thanks to the entire San Paolo group, a very genuine and passionate neuroclinical computational group. In particular, thanks to Gianvittorio Luria for all the inspiring discussions and wise suggestions.

Finally, but not in order of importance, I want to thank my reviewers for having read extremely carefully this work, for their questions and the several suggestions provided to improve the entire manuscript.

Contents

List of Figures

List of Tables

PART I

Introduction

1

Problems and Outline

Understanding neural activity both in its normal state and in presence of neurological disorders is one of the greatest challenges of this century [1]. A consistent part of the scientific community is putting intensive effort in the characterization of this complex organ, where the complexity arises from both its structural heterogeneity and the functional organization of its parts, giving origin to thoughts, actions, and more generally, perception, cognition, and behavior.

In the context of neurological disorders, such characterization is obtained through several medical tests, and requires a great expertise of trained clinicians, who invest a significant part of their time in visually inspecting different type of data, from long electrophysiological recordings to brain scans.

For the purpose of this story we invested our effort on signal processing and automatic methods aimed at interpreting these data. Given these premises, three years of investigation on this complex topic led us to the need of (i) defining standard protocols on the analysis of neural data, both in pre-processing and in the definition of relevant models, as this represents the starting point to generate reproducible results; (ii) developing some medical knowledge so to speak a common language with neurophysiologists; (iii) quantifying prior knowledge and clinical elements which are crucial for medical experts, but sometimes difficult to translate in mathematical terms.

In this work we explore a neurological disorder known as *focal epilepsy*. The focal epileptic condition is characterized by states of impaired consciousness, auras, and uncontrolled limb movements, which are known as *seizures*. Many causes can lead to the same diagnosis, among those brain trauma, lesions, tumors, and genetic factors. Differently from other types of epilepsy, in focal epilepsy the malfunctioning of the brain affects only a small portion of the brain. This portion is defined as the Epileptogenic Zone (EZ), or the minimum amount of brain area that being resected would lead to the absence of seizures. For drug resistant patients, ablation or surgical removal of the EZ may be the only solution for a regular daily life.

1 "The 21st century brain: Explaining, mending and manipulating the mind", Iain McClure

The EZ localization and the characterization of its neurophysiological activity through automatic tools is the purpose of this thesis. To this aim we will analyze electrophysiological characteristics that possibly describe and help in the identification of pathological zones. A not negligible aspect of this story is the high dimensionality of the dataset at hand. Indeed, we will deal with time series acquired at high sampling frequency (1 kHz), whose characterization by pure visual inspection can be extremely challenging, subjective and not reproducible, even when conducted by clinicians with long experience in the field. The automatization of the main clinical guidelines would be extremely useful to accelerate the analysis.

To approach this task, we heavily exploit signal processing methods and machine learning techniques. The former is essential to discard noise and preserve information in the signal, emphasizing some aspects of it, while the latter drives to the identification of the most important factors for the discrimination of the epileptogenic zones. We will leverage on regularized methods and feature extraction techniques to identify the main characteristics of the electrophysiological activity generated in the epileptogenic areas.

In this work, we focus on the analysis of invasive neural recordings during the *interictal stage*, which is the furthest time period from the epileptic seizure. At the behavioral level, the absence of any indication of the pathology would suggest the brain activity generated during this period to be the least informative to discriminate epileptogenic zones. Nonetheless, as we will observe through this work, the results presented here are encouraging and further investigations of the interictal period deserve to be studied. From a clinical perspective, the reader may ask why struggling for the design of methods which analyze the interictal stage, if the seizure onset time and other periods could be more meaningful to the localization of the epileptogenic zones.

Among many answers, one could be the search of a further comprehension of focal epilepsy as a chronic state rather than a short manifestation of pathological activity, the seizure. In the former scenario, the study of intracranial activity could be of great help for interpreting neurophysiological activity, and, at long term, be applicable in the analysis of non-invasive recordings. Non-invasive recordings have indeed a low spatial resolution, which is worsened by the presence of muscular artifacts, extremely likely during the seizure onset. In this perspective the analysis of interictal activity, and the transposition of any findings from invasive recordings to non-invasive ones would be of great help for clinical progress and improved diagnosis.

This work has been financially supported by the grant *"Advancing of non-invasive procedures for the support of early diagnosis of partial epilepsies"*, funded by Compagnia di San Paolo protocol 2017.AAI4513.U5101/SD/pv. The goal of this ongoing project is to provide support in the diagnosis of epileptogenic areas through the use of non invasive clinical tests, such as high density electroencephalography, during the childhood stage. This collaboration collects the effort of multiple departments, including DIBRIS (Dipartimento di Informatica, Bioingegneria, Robotica, ed Ingegneria dei Sistemi, Universitá degli

Studi di Genova), DIMA (DIpartimento di MAtematica, Universitá degli Studi di Genova), and Ospedale Pediatrico Istituto Giannina Gaslini, Genova, Italy.

1.1 Outline

The first part of the thesis gives a basic background on the main clinical aspects regarding the epileptic condition. In Chapter 2 we introduce the notion of epileptic condition, specifically the focal subtype. After a short description of the main medical tests performed in this clinical routine, we analyze in more detail invasive methodologies. We focus on invasive StereoElectroEncephaloGraphy (SEEG), which is the methodology used to record neurophysiological potential in deep brain structure. This technique has been central to our analysis, as all the applications described are designed on top of this data type. We further report the main clinical results related to the definition of biomarkers for the epileptic areas. We put major emphasis on patterns at high frequency, which have been pointed out as promising candidate biomarkers of the epileptic areas.

In Chapter 3 we focus on state-of-the-art tools in signal processing, in particular for the analysis of neural recordings. We introduce standard spectral analysis tools, as the Fourier transform; more sophisticated time-frequency characterizations of the signal, as the wavelet transform, and more recent data-adaptive representations, as dictionary learning. Then, we move to Machine Learning (ML) methods, in the supervised and unsupervised context. This represents a switch of context, which is showed to be useful as ML may represent an efficient tool for optimal feature extraction and could guide to efficient representation of the signal, driven by a learning task.

In Chapter 4 we report the clinical conditions for each patient used in the analysis.

Chapter 5 represents a starting point in the analysis of SEEG recordings. It consists of the application of clinical criteria and signal processing tools to extract some common clinical features from neural recordings. These methods are mainly based on the signal threshold, Fourier transform and wavelets. We do not respond to specific interpretation requirements, but rather to the first warm-up question. Is there any relevant information which may help in the discriminative task for the dataset at hand?

In Chapter 6 we describe a learning tool designed for a further step in signal interpretation, Multi Task Multiple Kernel Learning (MT-MKL). The method leverages on a multi-scale decomposition of the signal with the goal of selecting useful features to the discriminative task, common to epileptic patients. At the same time, MT-MKL generates a customized model for each patient, in which the prediction on the epileptogenicity of an area is performed based on similarities among new recordings and the learning recordings, used to build the model. Based on the neurophysiological activity the method identifies pathological zones by exploiting the correlation with pre-tagged areas. We descibe the results obtained from a first use of the learning pipeline, in which

the proportion of epileptogenic and physiological areas is almost balanced and the results from further analysis on the entire population. Throughout the chapter we will observe how, despite the good predictive performance, the method does not preserve its ability in selecting features. We formulate some hypotheses: i) a unique set of features might not be sufficient to capture the variability of a population of focal epileptic subjects; ii) the study of the entire time series might be not optimal, if the epileptic activity depends on rare pathological patterns.

Chapter 7 is the realization of the Occam's razor. In the light of the previous results, we look here for a further interpretation and the analysis of all pitfalls which may bias the results. At this point we try to quantify if automatic tools should be more focused small and rare time windows in the recordings, or rather the overall activity during the entire registration is more meaningful to the discriminative task. We will search through feature extraction and the imposition of sparse a priori the most determinant elements in the discriminative task. We will consider basic features which quantify the rare activity and others that represent the average behavior for a brain area. The last part of Chapter 7 will assume also higher clinical relevance, as corroborated by post-surgical outcome. It consists in the evaluation of the patients conditions after the surgical removal of the candidate epileptogenic zones, quantified through the Engel classification. We repeat indeed the analysis to the patients' subset correspondent to the best post surgical outcome. The results obtained here will guide us to the last part of the work, where we focus on smaller time windows.

In Chapter 8, which represents the final part of this work, we define a strategy for the search of pathological patterns of epileptic activity in the interictal stage. Given the previous results, and the need of translating this methodology to scalp measurements, we start by considering a restricted range of frequencies. We leverage here on open source toolbox. With the imposition of prior knowledge we report the results obtained from the attempt of selecting candidate epileptic waveforms.

PART II

Background

2

Focal Epilepsy and Diagnostic Tools

Here we provide a short characterization of epilepsy, with particular attention to focal epilepsy. We discuss the most common non-invasive diagnostic tools, with particular accent on methods relying on electrophysiological recordings. When these medical tests are not sufficient to precisely localize the epileptogenic tissue, more invasive medical tests are required. Thus we give a description of invasive intracranial measures, with a particular focus on StereoElectroncephalography. In the last part we present the main clinical results in this context and the main aspects in the electrophysiological signals which relate to the epileptogenic zones.

2.1 About Focal Epilepsy

Epilepsy is a neurological disorder mostly characterized by seizures, defined as *a transient disturbance of cerebral function caused by an abnormal neuronal discharge* [4]. Seizures manifest with different symptomatology which varies from involuntary movements involving part of the body, to loss of consciousness and convulsions. As reported by the World Health Organization [1], this condition affects 50 million people worldwide. Many factors can concur to epileptic symptoms, as lesions, brain trauma, brain tumors, and strokes, but in some cases epilepsy is determined by congenital and genetic factors, which lead to a malfunctioning of inhibitory mechanisms involved in signal propagation.

One of the main characterizations of epilepsy derives from the distinction among seizure types and lead to the definition of *general* and *focal or partial epilepsies*. In the former case, the malfunctioning depends on the entire brain, while in the latter only a restricted region of the brain is involved in the seizure generation. We restrict our analysis on this last category. As characterized by Aminoff et al. [4] partial seizures begin with motor and sensory phenomena and may involve clonic movements of limbs or of facial muscles. Physical symptoms can be pallor, flushing and sweating, while psychic symptoms lead to memory distortion, cognitive deficit, illusions, and hallucinations. In simple partial seizure, consciousness is preserved until the seizure discharge does not

1 https://www.who.int/news-room/fact-sheets/detail/epilepsy

spread to other brain areas than the focus, while in complex partial seizures consciousness, responsiveness, or memory are impaired.

In absence of lesions, the terminology used to define the brain regions related to the seizure generation and propagation discriminates between the area of primary organization of ictal discharge, named the *Seizure Onset Zone* (SOZ) and cortical areas involved in abnormal paroxysmal activity, the *irritative zone* (IZ) [58]. Throughout this thesis we use the more general definition of *Epileptogenic Zone* (EZ), which we refer to as pathological regions. The EZ is defined as the minimum amount of brain area such that its resection would lead to the absence of epileptic seizures.

A relevant portion (about 30%) of the epileptic patients does not respond positively to medical treatments [37]. For drug resistant focal epileptic patients a precise localization of the EZ followed by surgical intervention represents the only clinical scenario to improve their quality of life.

2.2 Non-invasive Diagnostic Exams

The diagnosis of focal epilepsy relies mostly on the manifestation of symptoms described above. The standard evaluation procedure relies on a set of non-invasive tests, which are listed in Table 1 [4]. These standard investigation techniques search both for abnormalities of the functional brain activity, through the analysis of electrophysiological recordings acquired with ElectroEncephaloGraphy (EEG), and for structural abnormalities, which are detected using Magnetic Resonance Imaging (MRI).

Table 1: Criteria for evaluation of new seizure disorder in a normal patient as in [4]

History	*medications and drug exposures*
General physical examinations	
Complete neurological examination	
Multiple EEG tests	
Brain MRI	*especially after the age of 25 years*

2.2.1 *Electroencephalography*

EEG represents one of the most reliable and well-established tool in the clinical routine. It is one of the first non-invasive tests conducted for the measurement of potentials generated by the brain activity. The first EEG recordings date back to 1924 and were performed by the German psychiatrist Hans Berger. In Figure 1 we report one of those registrations. The EEG records the collective activity of large ensembles of neurons also known as *neural populations* or *neural ensembles*.

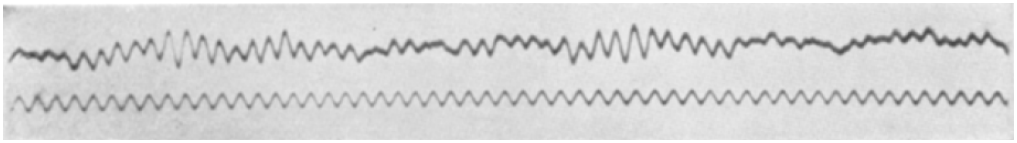

Figure 1: One of the first EEG recordings acquired from Hans Berger. In his preliminary studies Berger investigated both the physiological activity as well as epileptic alterations in electrophysiological recordings.

At its smallest scale the brain electrical activity arises from the generation of ionic currents in the neural cell, or neuron. The mechanism of neural activation for a single cell has been widely studied and characterized from the 50's, by the Hodgkin-Huxley model of the ionic mechanism underlying the generation and the propagation of synaptic potential in the giant squid axon [54].

Nonetheless, the signal generation and its propagation at larger scales are difficult and complex phenomena, as the collective behavior of a neural population gives rise to non-linear dynamical equations. These signals are also known as *post-synaptic potentials*. The post-synaptic potential from a neural ensemble with coherent orientation originates electrical variations which can be measured through scalp measure.

Modern EEG test, as a non-invasive procedure, relies on the acquisition of the brain potential through a set of sensors positioned on the scalp. The number of sensors, also called contacts or channels, changes depending on the EEG headset model, in a range from 19 to 256 sensors, respectively from the smallest to the highest spatial resolution. Each sensor records the electrical potential with respect to a reference signal. The EEG sensors show higher signal resolution for electrical contributions generated in the cortex, the closest brain region to the skull.

EEG gives a high temporal resolution signal, as the sampling frequencies may easily exceed the kHz. Despite that, as an external measure, it suffers from low spatial resolution, especially for activity generated in deep brain areas, also known as intracortical regions. Another factor which deteriorates the quality of the EEG signal is the presence of artifacts caused by muscular and respiratory activity, eye movements, and cardiac pace [109]. In our context, EEG is used as diagnostic tool for epilepsy by combining the resting state protocol to the analysis of sleep stages. It has been proved that the study of the two leads to 80% of correct diagnosis for focal epilepsy. In addition, EEG has been shown to be a highly specific and low sensitive instrument for the diagnostic task [112].

2.2.2 *Magnetic Resonance Imaging*

Magnetic Resonance Imaging (MRI) scans [115] reveal the structural characteristics of the brain and constitute the standard tool for excluding the presence of lesions or tumors.

Even if a plethora of MRI tests exist, all are based on the common principle of measuring the radio frequency signal emitted by atoms, due to changes in spins-orientations, as response to a magnetic field. In its basic realization (spin-echo MRI) the presence of hydrogen, and consequently of water molecules, is object of the measure. This allows to distinguish brain tissues with different water content, as fibers and grey matter. MRI can be effective in the diagnosis of epilepsy deriving from cortical dysplasia, or malformations involving the outer layers of cells in the cortex. Dysplasias manifest in three different types, may have genetic causes and well as depend on brain damage (type III). In the last case the test is sufficient to determine the pathological areas.

2.3 Invasive Diagnostic Exams

When non-invasive procedures fail in the localization of the EZ, further invasive medical tests are necessary. For patients diagnosed with focal epilepsy who do not respond to medical treatment, the removal or the ablation of the EZ through surgical intervention may be the only solution to improve their quality of life.

The presurgical assessment of the EZ through invasive procedure relies on intracranial measures aimed at localizing the pathological tissue. The localization must be precise and reliable, to reduce the amount of tissue to ablate, and consequent damages at the cognitive level.

Invasive techniques divide in ElectroCorticoGraphy (ECoG), where electrophysiological registrations are based on electrodes placed on the brain surface, and StereoElectroEncephaloGraphy (SEEG), which records the activity also from deeper brain structures.

We describe in detail the latter method, as the data analyzed in this thesis are acquired using SEEG.

2.3.1 *Stereoelectroencephalography*

SEEG has been first introduced by Talairach and Bancaud in 1974 [123]. In a more recent work, Cardinale et al. [22] give a precise characterization of the SEEG acquisitions protocol[2].

The SEEG is a surgical procedure which consists in the implantation of multiple filiform electrodes in the skull. The electrodes positions are determined from previous clinical assessments about areas potentially involved in the seizure generation. Through 3D-MRI scans and angiography, clinicians infer the precise mapping of blood vessels in these areas, so to exclude any possible complication during the intervention. The electrodes implantation takes place under general anesthesia, and a 3D Computer Tomography (CT) scan follows to check the correctness of the electrodes positions.

2 The data acquisition protocol for the results hereafter is the one described by Cardinale et al. [22], which is also the one adopted of our dataset, as collected in the same clinical center.

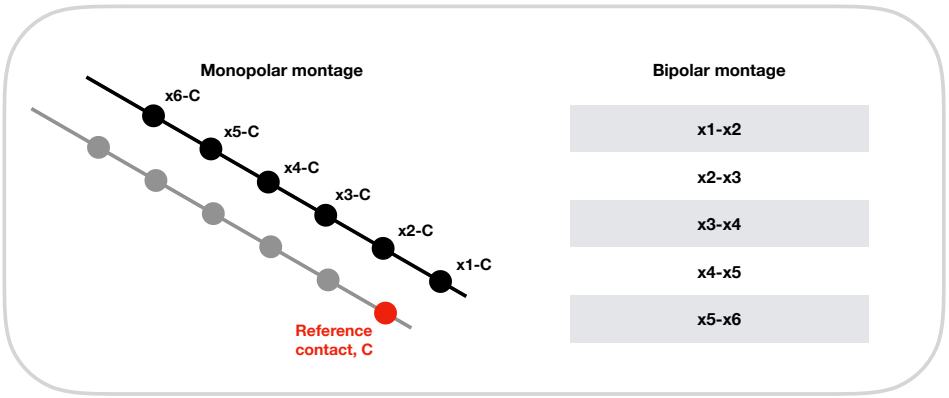

Figure 2: The raw measures are acquired in a monopolar setting, on the left. All the contacts from the same subject record the signal with respect to a unique reference. On the right, we convert the monopolar montage into a bipolar montage. This implies a spatial differentiation of neighbor recordings that are acquired on the same electrode.

The number of implanted electrodes and their positions are determined by the patients condition. In Chapter 4 we provide this information in detail. The contacts on each electrode register local field potential with a common reference in white matter. The community refers to this setting, in which recordings are acquired with respect to a unique reference, as *monopolar montage*. This case is reported on the left in Figure 2. The reference contact is highlighted in red.

It is a standard clinical procedure to refer the signal recorded at each contact with respect to the closest one on the same electrode. This is known as the *bipolar montage*, shown on the right of Figure 2. This is an approximation of the spatial gradient of the electrical potential, where strong signal variations correspond to high contributions from the recorded area.

As highlighted in the work of Mercier et al. [83] the monopolar montage does not get rid of contributions to the signal deriving from electrical volume conduction. The conversion to the bipolar setting is indeed considered a standard procedure to improve the spatial resolution of the signal.

Aside from a high spatial resolution, due to direct contact with cortical and deep brain structures, the SEEG recordings are characterized by a high temporal resolution, which depends on the sampling frequency, set typically above 1 kHz. A drawback of the high signal resolution is its spatial limitation. The measure is indeed operated on a small portion of the entire brain, and it does allow to study the activity with high spatial resolution from pre-surgical candidate regions only.

2.3.2 *Epileptic States and Surgical Intervention*

The localization of pathological areas still constitutes a challenging task even when supported by the use of SEEG acquisitions. The analysis of this long temporal recordings is typically performed through visual inspection by a team of medical experts, who label the signals. This procedure is extremely time consuming, prone to error, and subjective. The definition of the EZ relies on the analysis of multiple stages of the neural activity, at different electrophysiological states (e.g. sleep stages, resting state) and during different phases of the pathological conditions: interictal, preictal, ictal, and post ictal states[3].

The *interictal state* is the furthest from the epileptic seizure. During this period, the patient does not show pathological symptoms and conduce normal activities without any impairment.

The *preictal state* is defined as the period preceding a seizure. It can last from minutes to hours. Some patients attest some perceptive alterations e.g. aura or déjà-vu.

The *ictal state* is period when the seizure takes place. Impaired consciousness, loss of consciousness, convulsions, and motor symptoms are typical evidences of the epileptic seizure.

The *post ictal state* is the final phase after the seizure, which lasts about half an hour and is characterized by nausea, confusion, and headache.

Across different stages, the electrophysiological activity changes abruptly, especially in the transition from the preictal to the ictal stage. If the analysis of these phases is not revelatory for EZ localization, medical experts may rely on electrical stimulations of candidate epileptogenic areas through SEEG. This can support the clinicians in studying the neural response from each region and at the same time in evaluating possible cognitive damages deriving from its ablation.

As the epileptogenic areas have been identified, neurosurgeons proceed to surgical intervention. There are two possibilities, radio frequency thermocoaugulation or ablation of the pathological tissue. In the former case, the experts exploit the SEEG device to deliver currents causing a local temperature increase (78°-82° C) [17] in the candidate EZ. Ablation or physical removal of candidate EZ is historically the most consolidate surgical intervention, and follows to the SEEG investigation.

The surgical outcome is key in evaluating the performance in the EZ localization. The degree of success of the intervention is quantified through a scale of symptomatic signs of epilepsy which manifest after the surgery. One of the classical scale is the Engel classification, proposed publicly by Jerome Engel in 1992 [38]. The Engel scale discriminates the subjects in four categories. *Engel I* corresponds to seizure free subjects and represents the best surgical outcome; *Engel II* collects patients who rarely show disabling seizures; *Engel III* is relative to patients who manifest worthwhile improvement; *Engel IV* is for cases where no worthwhile improvement is shown. This classification suf-

3 https://www.epilepsycolorado.org/wp-content/uploads/2016/01/2-Types-of-Seizures.pdf

fers of subjectivity, as improvements assessment strongly depends on patients perception.

2.4 Biomarkers in Focal Epilepsy

The localization of EZ for drug resistant focal epileptic patients who do not show any signs of lesion is mostly based on clinical expertise, which consists in manually searching for pathological signature in the recordings. The subjectivity of this procedure has been highlighted from both the clinical and the computational communities [9], [66].

A vast literature aimed at defining biomarkers of epileptogenic areas in focal epilepsy exists. Several temporal short patterns and rhythms in the brain have been pointed out to be correlated to epileptic activity. Here, we propose a short review for the main results relative to the interictal stage.

2.4.1 *Interictal Spikes*

Interictal Spikes (ISs) are brief paroxysmal events of duration comprises between $100 - 300$ ms, and are retained a well-established signature of focal epilepsy [8, 82, 130]. An extensive characterization of ISs, on which we leverage later on, has been given in the work of de Curtis and Avanzini [32]. Pillai and collaborators gave a guideline to define electrophysiological patterns as ISs (see Table 2 in [100]). Typical ISs patterns are reported in Figure 3 from the work of Noebels and collaborators [92].

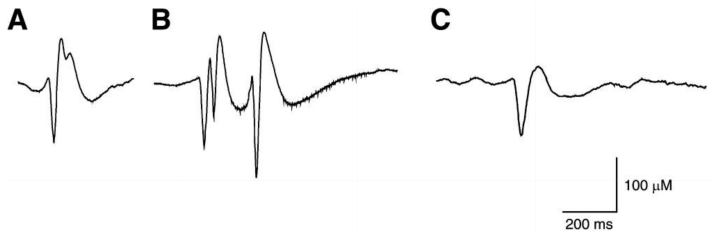

Figure 3: Figure from [92]. Different epileptic waveforms in human epilepsy from intracranial recordings are shown. A is an interictal spike, B is a group of interictal spikes, while C is a sharp wave from a lesional partial epilepsy.

In the last decades the mechanism underlying the ISs generation and their relation with seizures has been object of several studies.

De Curtis and Avanzini [32] advanced the hypothesis that IS may play an inhibitory role with respect to seizures. In 2002 Cohen and collaborators [25] studied at the cellular level the origin of interictal spikes in temporal lobe epilepsies, by analyzing epileptic neural tissue in vitro and inhibiting chemical receptors (GABA) at the cellular level. Staley and collaborators investigated the causal relation between ISs and the seizure onset for focal epilepsies derived from brain injuries, but excluded the causal implication between these phenomena in genetic epilepsy [121, 122]. In 2012 Wendling et al. analyzed

the capacity of microscopical and macroscopical models in the simulation of epilepticform discharges with particular attention to ISs, in both epileptic humans and animals [131]. In Karoly et al. [67] the authors investigated the periodicity between interictal spikes and seizure, throught a semi-automatic procedure for spike identification, based on matched template. Here the authors observed a correlation between the performance in seizure prediction performance and the amount of pre-ictal spikes.

2.4.2 *High Frequency Oscillations*

While there is common agreement on the importance of Interictal Spikes in focal epilepsy, more controversial results for other temporal patterns exist. A recent part of the literature analyzes other candidate epileptic signatures, called High Frequency Oscillations (HFOs). The main hypothesis, which justifies the efforts in this direction, is that the HFOs could be even more relevant than ISs in the focal area localization, due to their lower intensity and their locality [18, 60, 119, 137].

In [137] HFOs are defined as short patterns at high frequency (> 80 Hz), which have origin from the co-firing of small groups of interconnected principal cells.

One of the first criteria to define HFOs (Staba, 2002 [118]) requires that, a candidate HFO event must satisfy the following conditions a) the Root Mean Square (RMS) of the signal evaluated on a window of 3 ms $> 5\sigma(\text{RMS}_{\text{baseline}})$ (5 standard deviations above overall root mean square of the baseline), b) it must have a minimum duration of 6 ms, c) it must have at least 6 peaks above the 3σ of the rectified signal.

A further characterization divides HFOs in *ripples* (R), with typical frequency ranges between $80 - 250$ Hz, and *fast ripples* (FR), with typical frequencies in the range $250 - 500$ Hz. In FR this activity represents the collective effect of the activity generated from several populations of neurons, each firing at lower frequencies.

In Figure 4, a typical HFO pattern from the work of Fedele et al. [41] is reported. For HFOs, different clinical studies led to controversial results about their role as epileptic biomarkers. Concerning this, we report below some of the more recent results available from the literature.

The presence of FR with high probability links to seizure generation, as reported by Staba et al. [118]. In this work the authors report that HFOs are nonetheless involved in physiological activity, in the hippocampal regions, in the entorihnal cortex, and in the mesiotemporal lobe, where these are involved in memory formation and reactivation on previous experiences. Due to their shape, physiological and pathological HFOs can be distinguished when recorded using micro-electrodes, but this becomes challenging with macro-electrodes.

Presurgical localization of the epileptogenic zone is based on the identification of irritative areas, seizure onset zones (SOZ), epileptogenic lesions and functional deficit zones. It seems reasonable to add R and FR areas, keeping

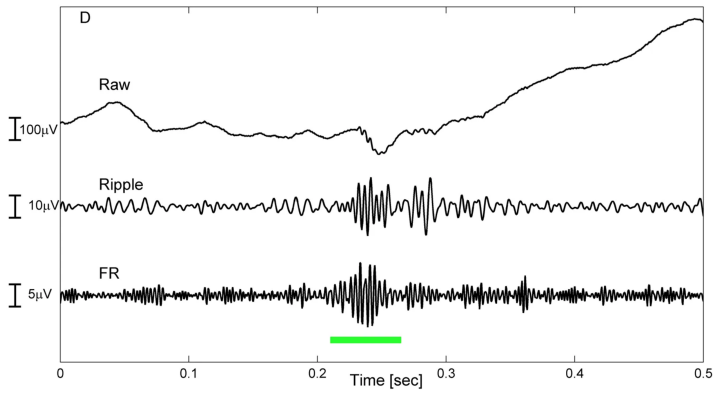

Figure 4: Co-occurrence of ripples and fast ripples in the trace, as the signal is high-passed respectively in the ranges $80 - 250$ Hz, $250 - 500$ Hz, [41].

in mind that the two mechanisms are related to different pathological mechanisms. The removal of cortex involved in the HFOs generation has been related to better post-surgical outcome than removing SOZ. The detectability of HFOs has been assessed even from non invasive test as MagnetoEncephaloGraphy (MEG) and EEG data [137].

An interesting meta-analysis of the role of HFOs and the surgical outcome is proposed in the work of Holler et al. [56]. To give a measure of the correlation of the HFOs with the EZ area the authors define first the *resection ratio* as

$$\text{resection ratio} = \frac{\#(\text{contacts w HFOs})^{\text{RA}}}{\#\text{ contacts w HFOs}} \tag{1}$$

The numerator is equivalent to the number of contacts containing HFOs in the Resected Area (RA), while the denominator corresponds to the number of contacts which registered HFOs for each subject. In this review the authors compare the resection ratio between seizure-free (sf) and not seizure-free patients (nsf) from 11 studies. In the following formula μ_x and V_x corresponds to the mean values and the variances for the the x population.

$$z = \hat{\mu}_{sf} - \hat{\mu}_{nsf}, \text{ mean difference of the two groups}$$
$$s^2 = \frac{1}{n_{sf}} V_{sf} + \frac{1}{n_{nsf}} V_{nsf}, \text{ inter-groups variance}$$
$$V_x = \frac{1}{n_x - 1} \cdot \sum_{i}^{n_x} (X_{x_i} - \hat{\mu}_x)^2, \text{ within-group variance}$$

In the last formula, X_{x_i} represents the individual resection ratio observed on patient i in group x. Citing the authors, one of the main results of this analysis shows that ripple resection ratio was higher in seizure free patients in 9 out of 10 studies, but as in 5 of 9 the positive confidence interval overlaps with zero, only 4 could be considered significant. For what concerns fast ripples the resection ratio is higher for the sf population in 5 out of 7 studies, but for 2 out of 5 the confidence interval overlaps with zero. In conclusion they assess that the statistical evidence of the relation between the resection ratio and seizure free outcome is quite poor and needs further investigations. Another

more recent work quantifies the HFOs predictive power of the epileptic tissue [41]. In this work Fedele et al. differentiate the contribute of R and FR, starting from the hypothesis that interictal HFOs are more specific than interictal spikes in the SOZ localization [60]. The authors make the point that most of studies reports a difference between mean HFO rates in SOZ electrodes and mean rates in non-SOZ electrodes. They made instead a classification of the cortex based on individual electrodes. In this study the authors consider 9 patients with mesial temporal lobe epilepsy and 11 with extra-temporal epilepsy, 13 seizure free patients after resection. Resorting to the previous work of [21] they detected ripples (median amplitude 27.4 μV peak to peak, interquartile range 15.0 μV), fast ripples (median amplitude 9.2 μV peak to peak, interquartile range 7.5 μV). They define the HFO contact as the one with HFO rate exceeding the 95% of the HFO distribution. R-FR areas are zones which show a co-occurrence of ripples and fast ripples.

	not seizure-free	seizure-free
HFO area not in RA	TP	FP
HFO area fully located in RA	FN	TN

Table 2: Definition of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) given for the patients considered in this study. The differentiation is made as the HFO area fully falls in the Resected Area (RA) or not.

The authors noted that, for the R-FR areas the 95 percentile criterion reaches a specificity ($\frac{\text{TN}}{\text{TN+FP}}$) of 100%, and 57% sensitivity ($\frac{\text{TP}}{\text{TP+FN}}$) (see Figure 1D [41]).

Last we cite one of the most recent results in terms of lack of significance of the correlation between HFO and the EZ areas [62]. The authors present here the results from a multi-center dataset (Freiburg, Montreal, Los Angeles), with a population of 52 subjects. The tagging of the HFOs during the sleep-activity at the interictal stage is partially guided by an automatic procedure [135], as the algorithm needs as input a baseline and some HFOs patterns. The procedure has been performed by two medical experts with good coherence in the evaluation of the patterns k-coefficient > 0.6. For what concerns the quantification of the overlap between HFOs areas and EZ, the authors characterized the contacts using two measures, as already done in Jacobs [61]. The first takes into account the rate of HFOs in resected and non resected areas, respectively RA and nRA

$$\text{ratio rate} = \frac{\sum_{i\in\{\text{RA}\}} \text{rate}_i - \sum_{j\in\{\text{nRA}\}} \text{rate}_j}{\sum_{k\in\{\text{all contacts}\}} \text{rate}_k}. \tag{2}$$

The last quantifies the ratio of contacts participating to high frequency events, regardless from the rate. This quantity partially relates to the resection ratio defined in [56], but does not allow a direct comparison.

$$\text{ratio contacts} = \frac{\#(\text{contacts w HFOs})^{\text{RA}} - \#(\text{contacts w HFOs})^{\text{nRA}}}{\#\text{ contacts w HFOs}}, \tag{3}$$

$$= \text{resection ratio} - \frac{\#(\text{contacts w HFOs})^{\text{nRA}}}{\#\text{ contacts w HFOs}}. \tag{4}$$

With $\#(\text{contacts w HFOs})^{\text{RA}}$, $\#(\text{contacts w HFOs})^{\text{nRA}}$ and # contacts w HFOs we denote respectively the number of contacts in the resected areas containing HFOs, the number of contacts containing HFOs in non resected areas and the number of contacts containing HFOs. The ratio contacts can assume here negative values (ratio contacts $\in [-1, 1]$) and it is more informative than the resection ratio as it discriminates the scenario i) no HFO events are registered in all contacts, from the scenario ii) HFOs events are present but only in non resected areas.

The results of this study are quite discouraging. The authors consider the post-surgical outcome for all patients to evaluate the relationship between the ratio rate and the ratio contacts with the Engel classification. The Spearman rank correlation coefficient $\rho \in [-1, 1]$ is used to quantify the monotonic relation between two ranked variables. Citing the authors *a deeper analysis of the predictive value in each center at the individual level revealed that HFOs did not reliably predict post-surgical outcome, with exception of the Los Angeles site.* The statistical results are shown in Table 1 [62].

In the light of the previous results, the result of Fedele et al. [41] is not in direct opposition to the more recent from Jacobs et al. [62], due to the different approach used in the definition of the HFO resection area. Indeed the TN rate of the former work can be traslate to

$$\text{specificity}_{\text{Fedele et al.}} = \frac{(\#\text{ contacts w HFOs})^{\text{RA}}}{\#\text{ contacts w HFOs}} = 1, \tag{5}$$

for seizure-free subjects (or Engel I subjects).

The TN definition reminds the resection ratio introduced in Eq. 1 with the peculiarity that to be defined # contacts w HFOs, a contact must meet strict conditions – the contact must exceed the 95% of the HFOs distribution. Due to this high threshold value, an abundant co-occurrence of ripples and fast ripples events in an area seems to be a *necessary but not sufficient condition to define an area as epileptic*, as the Engel I subjects with # contacts w HFO not in RA > 0 is null. The improvement in the resection ratio given by imposing a threshold to the amount of HFO events per contact so to consider it as a "contact w HFOs" had been proved to give smaller improvement to the resection ratio in the meta-analysis from Holler and collaborators, in Section 3.5 [56].

For further investigations, a larger population would be necessary to assess with high confidence that the HFO rate correlates to the EZ area. The analysis of this patterns presents moreover several problems, which make even harder their use in the clinical routine. These issues are mostly related to 1) the complexity of the analysis, e.g. filter artifacts may rise from ISs, giving similar

effects at high frequency 2) the definition of a baseline and HFOs templates require expertise and can be biased. In this regards the Cohen coefficient of agreement between visual and automatic marked data in [62] is equivalent to 0.6 in a range between $[0, 1]$.

2.4.3 *Alterations of Electrophysiological Rhythms*

The synchronization of neuron populations gives origin to oscillations and rhythmic patterns of electrical activity at a macroscopical level, which are widely visible from scalp measures as well. These rhythms, which are typical in the physiological brain, are known as *electrophysiological rhythms*. The community divides these oscillations in frequency bands of clinical interest, which are associated to different states. (i) δ rhythm, between $1 - 4$ Hz, is typically associated to the REM sleep stage. (ii) θ rhythm, $4 - 8$ Hz, is related to spatial navigation and memory. (iii) α rhythm, $8 - 13$ Hz, is emergent during the resting state activity, with closed eyes. (iv) β rhythm, in the range of $13 - 30$ Hz, is associated to active concentration and thinking. For (v) γ rhythm, $30 - 70$ Hz and (vi) high-γ between $70 - 90$ Hz, no common agreement on their functional role exists. Several works have pointed out altered electrophysiological rhythms as evidence of the epileptic status. In this concern we enumerate only a few among these. The work of Pyrowski et al. [103] highlights alterations of the α rhythm, which shows significative attenuation from the control population. The interval analysis reveals also an increase of the θ activity for the epileptic patients. Alterations of the θ rhythms have been observed also in [128], where the authors analyzed changes of this rhythm in relation with the network generated by areas showing HFOs. The work of Di Gennaro et al. [35] shows a correlation of Theta Intermittent Rhythmic Delta Activity (TIRDA) with the epileptic activity. TIRDA is defined here as a pattern of sinusoidal trains fo activity, in the range $1 - 3.5$ Hz with an amplitude comprises between $50 - 100\mu$V.

3

Data Representation and Learning

In this Chapter we will introduce different strategies for both fixed and adapted data representations. In the former case, the choice is independent from the dataset at hand, while the latter results in the search of an optimal representation, which translates in the definition of an optimization problem depending on the data at hand. We will introduce several optimization techniques developed in the context of signal representation and learning methods for solving supervised tasks. In this context the representation of the input data is guided by the need of predicting one or more variables. This connects to the goal of the work, as we tackle it as a binary classification problem, where we aim at discriminating neural recordings as acquired from epileptogenic or non epileptogenic zones. Learning to solve a specific task can potentially lead to the design of a useful representation of the dataset. We will conclude the chapter by discussing methods for unsupervised learning (e.g. clustering), as we will exploit them in the very last part of the thesis.

3.1 The Problem of Data Representation

Two main questions arise when discussing data representation. Why should I represent my data differently from the way they are given? What are the main advantages of this procedure?

There are tons of acceptable answers to these questions, ranging from physical needs *"My data are too large to fit in memory"*, to practical usage *"I do not want to scroll a long file of meaningless stuff before finding that small piece of information I desperately need"* or to the will of putting focus on some aspects in the data *"There are quantities I want to treat in different ways"*.

These basic motivations already reflect important necessities of people dealing with data analysis. The main driving forces are *selectivity*, *compression*, and *efficiency* [2, 50, 85]. Data representation improves data analysis by facilitating the comprehension of the phenomena under study and highlighting key aspects related to their statistical properties. Even when not explicitly stated, there is typically an interpretation need which guides to the choice of a specific representation.

Automatic learning methods have a consistent intersection to the data representation problem. Starting from the data, they leverage on automatic approaches to solve a task. Learning methods are based on the formulation of a mathematical cost function and the use of optimization techniques to find the optimal model to fit the data.

Throughout the chapter we will notice that learning and data representation techniques are strictly intertwined. Historically, learning methods sat on top of pre-made and ad hoc representations depending on the data at hand and guided by domain expertise. Through more recent methods the representation could have been redefined, based on regularization and feature extraction techniques. In the last years, due to an increasing complexity of the models used for learning task, data-driven representations emerge as a side effect of prediction.

To make the picture clearer, we report a diagram in Figure 5, which shows the main aspects we will review in this chapter. We refer the reader to books and papers which we cite throughout the analysis for an exhaustive explanation of the introduced topics.

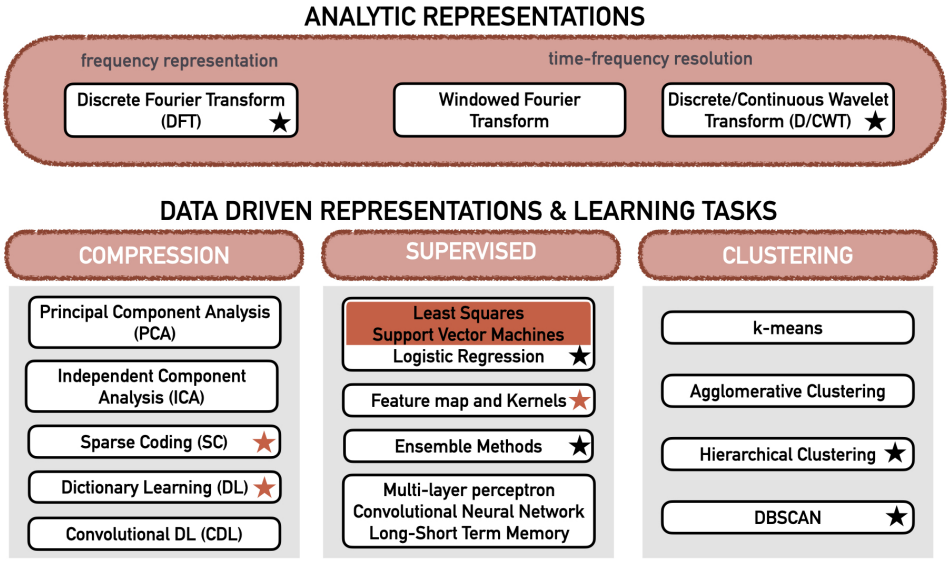

Figure 5: Schematic representation of the main topics covered in this Chapter. At the top, standard tools for time series representation: Fourier and time-frequency analysis. At the bottom, most popular learning methods where the representation responds to specific tasks as compression, supervised learning, and clustering. Topics denoted with a black star are methods which we made us of, while the ones with red star are areas to which we made a contribution presented in the thesis.

The problem of data representation is particularly important for the analysis of time series. Indeed considering single time points is in general not a good choice for understanding the data, while analyzing the collective behavior of the time series may be. The understanding of phenomena underlying time series is typically difficult due to the high dimensionality of the data, unknown dynamics, and plausible causal relations among variables, for the multivariate

case. Note that, even if causality is a topic of extreme interest for the analysis of our data, we will not approach it in the thesis.

Starting from the top of Figure 5 we introduce hard-coded data representation methods, which include spectral analysis, or Fourier analysis, and more sophisticated tools for time-frequency representations for the signal, as windowed Fourier transforms and wavelet transforms. All these representations share the common peculiarity of being invertible transformations, implying no information loss with respect to the original signal.

The representation may also be guided by the aim of responding to a specific task, we will see how learning methods and optimization play a major role here. The topics in the bottom part of Figure 5 mostly cover these aspects. We identified three main goals where representations and learning are strictly interconnected: *compressed representations*, *supervised tasks*, and *clustering*. We will describe some techniques for each of the topic, putting emphasis on the ones we actually use, or to which we contribute. Given the question in mind and the expertise about the domain, we analyze also how to inject this information in the solution, which is usually referred to as *prior knowledge imposition*. This concept will occur frequently throughout the chapter, and we will explore some of the strategies used to impose a priori to the solution of an optimization problem.

3.2 Time Series Representations

Time series and their representation are the main protagonists of this thesis. In particular we will deal with intracranial variations of electrical potential. Even though these quantities are continuous in time, the observation of a signal is affected by discretization, where the temporal resolution is imposed by the sampling frequency of our acquisition system. In the next sections we denote through $S(t)$ a generic time series. The transformations presented in the following have been formalized in both the continuous and discrete temporal domains. Here we present their continuous version, as the focus in not discussing their numerical implementation but rather giving an intuition about their effect on the data representation. It is worth noticing that these transformations are invertible, implying no information loss neither compression, while may plausibly improve the interpretation of the signal.

3.2.1 *Spectral Analysis*

The standard approach in spectral analysis consists in finding a decomposition of the signal onto patterns of periodic behavior. As introduced before, meaningful information in an electrophysiological recording can be present at different level, or frequency bands.

Historically Fourier analysis provides the most well-established tool for the analysis of time series, as it decomposes the original signal into a sum of sinusoidal functions of different frequencies and phases.

Given a time series $S(t) \in L^1(\mathbb{R})$, the space of integrable functions, the Fourier transform $\mathcal{F}$ maps a signal $S(t)$ from the temporal to the new representation in the frequency domain $\hat{S}(f)$ as follows

$$(\mathcal{F}S)(f) = \int_{\mathbb{R}} dt S(t) \exp\left[-i2\pi ft\right] = \hat{S}(f). \tag{6}$$

by projecting S on an infinite basis of sinusoidal functions. The numerical realization of the Fourier transform relies on a discrete version for the continuous Fourier transform, and it is known as Discrete Fourier Transform (DFT) [28]. In the thesis we will heavily use this tool, which represents a gold standard for time series analysis.

3.2.2 *Time Localization*

The projection over a periodic signal presents some limitations. The Fourier decomposition in Eq. 6 is not optimal to represent short transient in the time series as this is rather a global transformation, which takes into account the entire signal in the temporal domain. There are other time series representations which overcome this issue as the windowed Fourier Transform and the Wavelet Transform.

3.2.2.1 *Windowed Fourier Transform*

The windowed Fourier transform provides a temporal localization by defining a window function g, typically smooth with compact support which selectively focuses on a signal portion in the temporal domain. The transform is defined as

$$(T^{\text{window}}S)(f,\tau) = \int d\xi S(\xi)\; g(\xi - \tau) \exp\left[-if\xi\right]. \tag{7}$$

T^{window} denotes the windowed Fourier transform over the signal S, which is defined on both the temporal and frequency domains. The function g is the smooth window convolved with the original signal. The exponential term in the integral has the same role as in the Fourier transform, and it is a sinusoidal function whose frequency is specified by f. The size of the window function represents the trade-off between temporal and frequency resolution. Short windows in the temporal domain will localize transients with high temporal resolution, leading nonetheless to a worse description in frequency. Indeed as we reduce the width of the window, the transformation gets further from being a pure oscillation, for decreasing values of f in Eq. 7. In Figure 6 we show an example of the basis $g(\xi - \tau)\exp\left[-if\xi\right]$ in Eq. 7, for different frequency values, $f = 2$ Hz, on the left, and $f = 10$ Hz, on the right. The window support is independent from the frequency value, and this aspect, as highlighted by Daubechies, may represent a limitation as it does not allow an optimal localization for both sharp and wide transients in the signal [31].

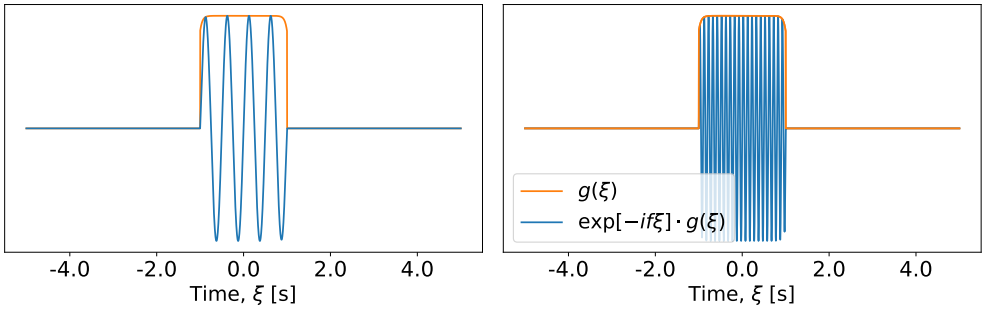

Figure 6: Two examples of $g(\xi - \tau)\exp[-if\xi]$ for a window of fixed length, orange curve. The window g has been constructed using functions with exponential decay. There is a gain in terms of locality with respect to the Fourier transformation, but the results could be not optimal as the frequency increases. The window width does not change as the frequency increases, as shown on the right.

3.2.2.2 *Wavelet Transform*

In this regard we introduce a last family of transformations, known as wavelet transforms, which offer a valid solution to the localization issue, as they allow to get a good localization of short as wide transients in the signal. Differently from the previous methods, the wavelet transform requires the definition of a function Ψ, called mother wavelet, which is not sinusoidal, onto which we project the data. In Figure 7 we provide an example of mother wavelet for different scaling parameters. To be more precise two types of wavelet trans-

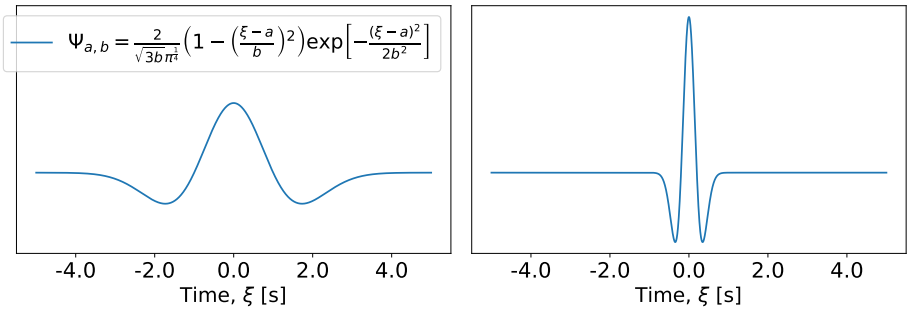

Figure 7: Example of a mother wavelet (Mexican hat), whose analytical form is defined in the legend. The two examples reported here are two versions of the same function at different scales. The scale parameter, related to the frequency, corresponds to b. Differently from what we have seen before for the windowed Fourier transform, a change in the sharpness of the function affects its support.

form exist: the Continuous Wavelet Transform (CWT) and the Discrete Wavelet Transform (DWT).

CONTINUOUS WAVELET TRANSFORM The CWT requires the choice of a function $\Psi_{a,b}$, also known as mother wavelet which is parametrized by $a \in \mathbb{R}$,

the scaling parameter, and $b \in \mathbb{R}$, the translation parameter. The former allows to dilate or compress the function changing at the same time its support.

$$(\mathcal{W}^{\text{CWT}}S)(a,b) = \frac{1}{\sqrt{a}} \int d\xi S(\xi)\ \Psi_{a,b}(\xi) = \frac{1}{\sqrt{a}} \int d\xi S(\xi)\ \Psi\left(\frac{\xi - b}{a}\right) \tag{8}$$

Besides the requirement over Ψ of square integrability in $\mathbb{R}$, the *admissibility condition* is necessary to define the wavelet transform at hand as invertible.

$$C_\Psi = 2\pi \int d\xi |\xi|^{-1} |\hat{\Psi}(\xi)| < +\infty, \text{ admissibility condition.} \tag{9}$$

DISCRETE WAVELET TRANSFORM Differently from the CWT, in the DWT the parameters (a,b) assume discrete values. The discrete wavelet is thus defined, at the j-th scale, as

$$\Psi_{j,k} = \frac{1}{\sqrt{2^j}} \Psi\left(\frac{t - k2^j}{2^j}\right), \text{ with } j,k \in \mathbb{Z}. \tag{10}$$

The result of the transformation is

$$(\mathcal{W}^{\text{DWT}}S)(j,k) = \frac{1}{\sqrt{2^j}} \int d\xi S(\xi)\ \Psi_{j,k}(\xi) = \frac{1}{\sqrt{2^j}} \int d\xi S(\xi)\ \Psi\left(\frac{\xi - k2^j}{2^j}\right). \tag{11}$$

The transform gives the result of convolution at discrete step, for a fixed scale j. The convolution is indeed computed at position $[0, 2^j, 2^{2j}, \ldots, 2^M]$, with $M = \frac{\log_2 N}{j}$.

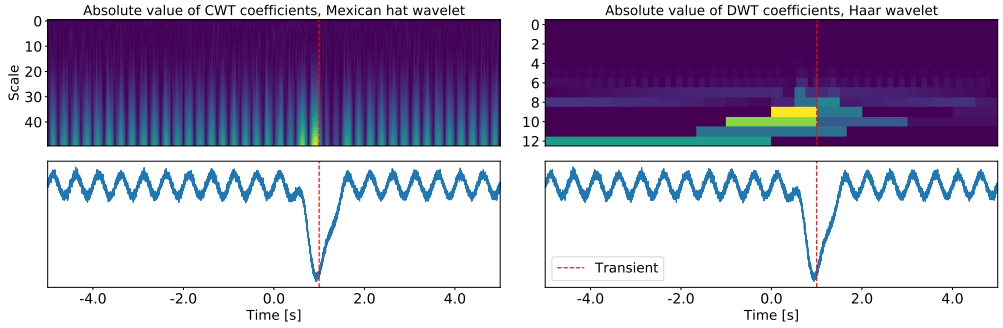

Figure 8: Wavelet coefficients of a signal S showing a transient $S(t) = \sin(\omega_0 t) + \exp[(t - t_0)^2/(2\sigma^2)] + \mathcal{N}(0, \sigma_n)$. On the left: the CWT with Mexican hat mother wavelet. The output is a matrix of coefficients whose dimensions are (#s $\times$ N), with #s choice of the user. On the right: DWT coefficients, with Haar mother wavelet.

In Figure 8 we report the absolute value of the wavelet coefficients obtained from a non periodic signal. The wavelet coefficients are able to catch the transient with a good temporal resolution. The peak occurs in presence of the dashed red line. The two transformations differ in the output dimensions. On the left, the array of coefficients of the CWT has the same length across scales, due to the continuous translation in the temporal axis. The number of DWT coefficients instead increases as the scale decreases. For the DWT, as shown in

Formula 10, the translation is discretized. Due to the geometric sequence of ratio 2 for the scaling parameter this wavelet is also defined as *dyadic wavelet transform*.

The choice of the mother wavelet is a key aspect for the output representation. If we are searching for a specific waveform in the signal, the mother wavelet should resemble the desired waveform, so to localize at best the pattern. A typical choice is the Morlet wavelet, widely adopted in neuroscience [20, 75] and implemented in standard tools for the analysis of electrophysiological data[1].

3.3 Learning Methods for Compressed Representations

So far we introduced lossless data representation methods for time series. Nonetheless, data representation may become extremely challenging for high-dimensional data and may be not optimal in terms of data storage. In this regard, compressive methods play a crucial role.

In practice this requirement entails a dimensionality reduction, which may lead to truncation or, for more advanced scenarios, to the imposition of constraints on the resulting representation. Advanced techniques for data compression take the form of optimization problems. In this sense, an optimal compression should lead to small information loss, while reducing the dimensionality of the data. Compression, as solution of an optimization problem, represents the optimal representation learned from the data, as it is guided by the data themselves.

For what concerns time series representation, compressed versions of the wavelet transforms derive from coefficients thresholding [23], coefficients shrinkage [36], and high frequencies removal. Several approximation schemes, whose choice depends on the hypothesis of regularity of the input signal, are discussed in Chapter 9 [79]. As we have seen, the Fourier or wavelet transforms are representation tools independent on the specific time series at hand but rather guided by previous assumptions on the phenomena under study.

In this chapter we will further explore optimization schemes whose outcome are both the set of generators used to project the data and the set of coefficients, both shaped by the dataset.

We introduce some further notation. As some of the methods described below are not specifically designed for time series, let $x \in \mathbb{R}^p$ denote a generic observation of p scalar quantities, which can be viewed as one among N observations. Those are collected in the dataset, $X \in \mathbb{R}^{N \times p}$. We explore next several methods for dimensionality reduction. Those explicitly designed for time series will leverage on the previous notation for signal, S.

1 Last access: May 4th, 2020https://neuroimage.usc.edu/brainstorm/Tutorials/TimeFrequency

3.3.1 *Principal Component Analysis*

Principal Component Analysis (PCA) is the most basic version of matrix factorization [97]. PCA is a standard method for dimensionality reduction in many fields, including image analysis [76], signal processing [111] and data visualization [65].

In the hypothesis of possibly correlated measures, the method returns a decomposition into orthogonal components. The first component combines the original features and denotes the direction of maximal variation for the given set of observations [1].

Each column of the input data matrix X must be centered so to have zero mean. The direction of maximum variance consists in the solution to the following optimization problem

$$w^{(1)} = \arg\max_{\|w\|=1} \left\{ \|Xw\|^2 \right\} \tag{12}$$

Given the constraint of unit norm, the problem is equivalent to

$$w^{(1)} = \arg\max_{\|w\|=1} \left\{ \frac{w^T X^T X w}{w^T w} \right\}.$$

The solution corresponds to the largest eigenvalue for the matrix $X^T X$, with $w^{(1)}$ the corresponding eigenvector. Once we compute this quantity, the first component is equivalent to $t^{(1)} = X \cdot w^{(1)}$. The procedure is repeated iteratively for the other components, with the only difference that the matrix considered at the k-th iteration becomes $\tilde{X} \leftarrow X - \sum_{s=1}^{k-1} X w^{(s)} \left(w^{(s)} \right)^T$, which is the residual of the original matrix, given the components which we already computed. Then, the estimation of the k-th component can be formulated as

$$w^{(k)} = \arg\max_{\|w\|=1} \left\{ \frac{w^T \tilde{X}^T \tilde{X} w}{w^T w} \right\}.$$

The full principal component decomposition of X can then be written as

$$T = XW \tag{13}$$

where W is the square matrix whose columns are the ordered eigenvectors for $X^T X$. This result can be obtained equivalently through matrix diagonalization, using the algorithm for Singular Value Decomposition (SVD) [44]. A reduction of the output dimensionality can result from eigenvalue truncation (and corresponding eigenvectors), which corresponds to the removal of directions with the smallest variance. Typically this is achieved by fixing a level of variance we want to preserve from the original data, and by keeping the corresponding number of PCA components.

3.3.2 *Independent Component Analysis*

PCA is based on a second order statistics, or variance. Differently, Independent Component Analysis (ICA) exploits higher order statistics. ICA separates a

multivariate signal as the sum of different sources, which are assumed to be statistically independent. Given the dataset X, ICA aims at finding

$$X = AZ \tag{14}$$

with Z, matrix of independent sources, where each row is the realization of random process and A is the mixing matrix. As in PCA, the ICA algorithm [57] assumes zero mean input, which can be easily achieved by re-centering the data. One among the most popular choice to implement ICA consists in the search of the decomposition which maximizes the non-Gaussianity of the sources. This choice is based on the result of the Central Limit Theorem, which states that, as the number of independent random variables N increases, the probability distribution for their sum converges to a Gaussian distribution. Given the hypothesis of independent non-Gaussian random variable, the idea is finding iteratively the components which give the most non Gaussian set of signals. For further details we refer to [57], Section 4. Non-Gaussianity is not the only option to achieve the ICA decomposition. For information based criteria we refer again to [57].

3.3.3 *Adaptive Matrix Factorization*

With adaptive matrix factorization techniques we denote a family of methods which share the common characteristic of factorizing a data matrix in the product of matrices

$$X \sim CD, \tag{15}$$

with $\dim(C) = N \times k$, $\dim(D) = k \times p$. The exact factorization is possible if the condition $\text{rank}(X) \leq k$ holds true [116]. Despite PCA and ICA fall in the family of matrix factorization techniques, the interest in adaptive matrix factorization methods arises nonetheless from the necessity of approximating high dimensional matrices of low rank as the product of smaller matrices. In presence of small dimensional representations, such that $\text{rank}(CD) < k$ with k user parameter and $k \ll \text{rank}(X)$, it is possible to talk about *low rank matrix factorization*. In particular with adaptive matrix factorization we denote Sparse Coding (SC) and Dictionary Learning (DL), two popular matrix factorization methods. Their adaptivity is given by explicit requirements of interpretability, sparsity, or compressibility imposed through specific cost functions.

In [125] we presented a library to optimize this type of problems. Hereafter, we give a short characterization of SC and DL. Feel free to refer to [125] for further details about their implementation.

3.3.3.1 *Sparse Coding*

Given X, sparse coding decomposes the samples on a set of k vectors. Let $D \in \mathbb{R}^{k \times p}$ be a fixed *dictionary*. The problem does not simply imply the projection of the original data on D, as prior knowledge can be imposed on the output representation. The aim of sparse coding is finding the best C, matrix

of coefficients, such that $X \sim CD$, with $C \in \mathbb{R}^{N \times k}$, where we may constrain the solution C. The optimization problem formulates as follows

$$C^* = \arg\min_C \{\|X - CD\| + \lambda\Psi(C)\}, \tag{16}$$

where, on the right side of the equivalence, the left term represents the distance between the approximation and the true data. We do not specify the metric, which is a user choice. The last term on the right represents the constraint, expressed through the functional Ψ. Typical constraints are: sparsity, achieved through L^0 or L^1 norms, shrinkage, achieved using L^2 as also a measure to avoid correlation, together with the combination of the L_1 and L_2 norms. Depending on the choice of the norm, the problem is strictly convex and admits a unique solution, e.g. in the L^2 case. The parameter λ represents a trade-off between the approximation of the input data and the weight of our constraint. Its choice is a crucial part of the optimization problem. Among the possible values it can assume (which are fixed to be in a finite vector, or can be the result of a random sampling strategy), it is chosen as the one that give rise to the more robust results and best approximations of the original data.

3.3.3.2 *Dictionary Learning*

DL is another matrix factorization method where both the coefficients C and the dictionary D are learnt from data. Given the same hypothesis of SC, or $X \sim CD$, the optimal C^* and D^* matrices are results of the following minimization problem

$$(C^*, D^*) = \arg\min_{C,D} \{\|X - CD\| + \lambda_1\Psi(C) + \lambda_2\Phi(D)\}. \tag{17}$$

The matrices dimensions $C \in \mathbb{R}^{N \times k}$ and $D \in \mathbb{R}^{k \times p}$ depend on k, a free parameter of the method which fixes the number of atoms in the dictionary. The parameters λ_1 and λ_2 weight the importance of prior knowledge, which is imposed through the functionals Ψ and Φ respectively.

For both SC and DL, the flexibility given by the choice of the fitting data term and the constraint terms is highly desirable.

Typical choices are L^1 norm terms under the assumption of sparse solutions (Chapter 7.6 [51]), L^2 norm to shrink and smooth coefficients [116], L^∞ term to penalize the presence of vector components with high amplitude [126, 127].

Other constraints may rise by requiring a low rank solution, which we can imposed through a constrain on the rank of the matrices or through a small value of $k \ll \min(N, p)$. These constraints can be imposed on both the matrices C and D.

Problem 17 is not jointly convex in the variables C, D. The strategy proposed for the search of optimal parameters in [125] has been proximal alternating gradient descent [14]. Following this criterion, the convergence to a local minima, under the choice of a reasonable learning step, is guaranteed.

3.3.4 *Convolutional Dictionary Learning and Sparse Coding*

All the compressive strategies introduced above take into account the entire input dataset, which, in case of time series, is the entire period of acquisition. Some recent methods used the concept of matrix decomposition to provide adaptive representations for time series. Those go under the name of Convolutional Sparse Coding (CSC) and Convolutional Dictionary Learning (CDL). Similarly to SC and DL the idea is to decompose the original time series S using patterns of smaller support than the entire signal length. This decomposition is obtained through the convolution of patterns and coefficients, which are alternatively optimized. We report the clear formulation given in La Tour et al. [71] which gives an overview on the available CDL implementations.

3.3.4.1 *Univariate CDL*

Given a set of one dimensional signals $\{S^n\}_{n=1}^N \subset \mathbb{R}^T$ of length T, the optimization problem corresponds to

$$\min_{\{d_k\}_k, \{z_k^n\}_{k,n}} \left\{ \sum_{n=1}^{N} \frac{1}{2} \left\| S^n - \sum_{k=1}^{K} z_k^n * d_k \right\|_2^2 + \lambda \sum_{k=1}^{N} \|z_k^n\|_1 \right\}, \tag{18}$$

$$\text{subject to } \|d_k\|_2^2 \leq 1, \text{ and } z_k^n \geq 0. \tag{19}$$

Here λ is the regularization parameter, $\{d_k\}_{k=1}^K$ is the dictionary of patterns of support L, and $\{z_k^n\} \subset \mathbb{R}^{T-L+1}$ are the coefficients related to the activation for the k-th pattern and the n-th time series. One of the first implementations of this method has been proposed in [49] in the context of audio classification.

3.3.4.2 *Multivariate CDL*

In the multivariate scenario we deal with the evolution of multiple variables. In this case each sample is a family of C time series, each of length T, for a total of N samples. We denote the generic j-th multivariate time series as $(MS)^j \in \mathbb{R}^{C \times T}$.

$$\min_{\{D_k\}_k, \{z_k^n\}_{k,n}} \left\{ \sum_{n=1}^{N} \frac{1}{2} \left\| MS^n - \sum_{k=1}^{K} z_k^n * D_k \right\|_2^2 + \lambda \sum_{k=1}^{N} \|z_k^n\|_1 \right\}, \tag{20}$$

$$\text{subject to } \|D_k\|_2^2 \leq 1, \text{ and } z_k^n \geq 0. \tag{21}$$

The k-the element of the dictionary is $D_k \in \mathbb{R}^{C \times L}$ and the activations coefficients, are, as for the univariate case $z_k^n \in \mathbb{R}^{T-L+1}$. This approach has also been implemented for the analysis of multivariate neurophysiological recordings as EEG signals, see Barthelemy et al. [10].

3.3.4.3 *Multivariate with Rank-1 CDL*

The method proposed in La Tour et al. [71] has been explicitly designed for neural recordings, in particular EEG and MEG signals. Here, leveraging on

the assumption that the real dimensionality of the multivariate time series may be much smaller than the number of recordings, the authors define a low rank convolutional dictionary learning approach. The multivariate-time series is approximated by a single temporal pattern, generated in a region of the brain, that propagates through all brain regions. The amplitude of the patterns is modulated by a vector of weights, of the same dimension of the number of recordings. The dictionary $D_k \in \mathbb{R}^{C \times L}$ of the Multivariate CDL is decomposed as $D_k = u_k v_k^T$, product of $u_k \in \mathbb{R}^{C \times 1}$ the spatial vector and $v_k \in \mathbb{R}^{L \times 1}$ the temporal vector. Again the activations for each element k of the dictionary are encoded by the coefficient z_k^n.

3.4 Learning Methods for Supervised Tasks

As we have introduced before, optimization strategies solve many different tasks, not necessarily related to compression. A substantial part of machine learning literature was developed to answer predictive challenges. Image recognition, audio classification, sentiment analysis, and decision systems are few examples of predictive challenges tackled by machine learning techniques.

The traditional learning paradigm for supervised tasks is such that, given a set of data $(X_i, Y_i)_{i=1}^N$ we aim at finding a relation f which maps our input data in the output quantity. The generic i-th sample, $(X_i) \in \mathbb{R}^p$, is an array of p features, while the correspondent output value Y_i can be equivalently an array or a scalar quantity. The relation f is inferred from the dataset at hand, which constitutes the *learning set*.

In case of discrete output values Y representing a category we refer to the problem as a *classification problem*. In the case where Y is a continuous quantity, the learning task takes the name of *regression problem*.

In both scenarios, given the tuple (X_i, Y_i) corresponding to the generic i-th observation, and $\mathcal{X} \times \mathcal{Y}$ as the space of the possible realizations for (X, Y) variables, we make the hypothesis on an underlying joint probability distribution $p(X, Y)$. The joint probability models different sources of uncertainty, depending on both the input and the output variables. Under the main assumption that the probability distribution of X is independent from Y, the joint probability is equivalent to

$$p(X, Y) = p_{\mathcal{X}}(X) p(Y|X).$$

The distribution $p(Y|X)$ models the noise in the output, meaning that given a fixed X there is not a determined relation which maps it with probability 1 to a unique Y value. Some techniques (e.g. Bayesian methods) also deal with the estimation of prior probability distributions, but these approaches typically require strong assumptions on the process generating the input data. As already mentioned, from a set of data, our learning task will be limited to the search of a predictive model, that involves the conditional probability only.

Pivotal ingredients of this learning paradigm are: (i) the *dataset*, which should be as large as possible in terms of number of N samples. The quantity of sample needed to build a model is a critical aspect in order to get generalizable

results, but can be out of control if the experiments have been not explicitly designed a priori. Moreover scientists must fix carefully (ii) the *hypothesis class*, a family of functions which may approximate the input-output relation. This can be an extremely wide set and differently from standard parametrical methods, the choice of the best function comes as the consequence of the learning process, where the output model and its complexity are results of an optimization procedure and depends on the dataset [3]. In the following section we report examples of basic hypothesis class, e.g. linear models, as well as more flexible non linear models. The third key aspect of learning is the definition of (iii) an *optimization problem*, implying the choice of a loss function $\mathcal{L}$. The functional $\mathcal{L}$, given X_i, quantifies the error made by the model in predicting Y_i, and it maps $\mathcal{L} : (Y; X, f) \to [0, +\infty)$. It typically consists in the sum of the loss function for all the points used for model inference, as reported in Formula (22)

$$\mathcal{L}(Y; f, X) = \sum_{i=1}^{N} \mathcal{L}(Y_i; f, X_i). \tag{22}$$

The choice of the loss function is a crucial aspect, which has drastic effect on the model performance. In Table 3 we report some typical loss functions for classification and regression. Typically, the hypothesis class corresponds

Binary classification	Regression
Least square $\|1 - Yf(X)\|$	Least square $\|Y - f(X)\|$
Hinge loss $\|1 - Yf(X)\|_+$	L^1 norm $\|Y - f(X)\|$
Logistic $\log(1 + \exp[-Yf(X)])$	ε-insensitive $\|Y - f(X)\|_\varepsilon$

Table 3: Typical loss functions for binary classification (left), and for regression problems (right). The two functionals can be easily generalized to the multi-category/multi-regression cases.

to the class of linear functions. In the regression case, the function can be written as $f(X) = X\beta + b$, while in binary classification it may correspond to $f(X) = \text{sgn}(X\beta + b)$. Given the training tuple (X, y) this translates to the search of the tuple (β, b) which gives the best prediction $f(X)$, or the least distant from the output.

Nonetheless linear models can be limited in terms of predictive capacity in presence of more complex underlying dependence of the output from input variables. This drawback can be overcome through non-linear models, which may be more complex but more flexible, in term of data fitting. Among those we make a distinction between *kernel methods* and *non-linear methods*.

3.4.1 *Kernel Methods*

With kernel methods we denote a family of methods whose strategy is to represent the input data in a typically higher dimensional space. The mapping

to higher dimensional space is obtained by fixing a *feature map*, $\phi : \mathbb{R}^p \to \mathcal{H}_k$, where $\mathcal{H}_k$ denotes a Reproducing Kernel Hilbert Space (RKHS) [7]. This space has the following properties

1. as subspace of the Hilbert space, it is endowed of distance and norm;
2. it is the space of functions which, given an element in $\mathbb{R}^p$, with p feature space, map to $\mathbb{R}$

$$\mathcal{H}_k = \{f : f \in \mathbb{R}^p \to \mathbb{R}\};$$

3. as we define an evaluation functional T_x, with $T_x : \mathcal{H}_k \to \mathbb{R}$,

$$T_x(f) = f(x)$$

this is continuous.

Given a feature mapping ϕ in a RKHS, the non-linear function in the original space can be written as

$$f(\cdot) = \phi(\cdot)\beta + b, \tag{23}$$

with an additional threshold in presence of a classification task. The function is non-linear in the original input feature space, but in the tuple (β, b).

To give an intuition we report an example about the beneficial usage of a feature map, in Figure 9. The image represents a binary classification problem, where our input samples are originally vectors in a $\mathbb{R}^2$ space, on the left. Each sample is represented as a dot; the two classes, encoded by two different colors, are non-linearly separable in the original space. Using a polynomial

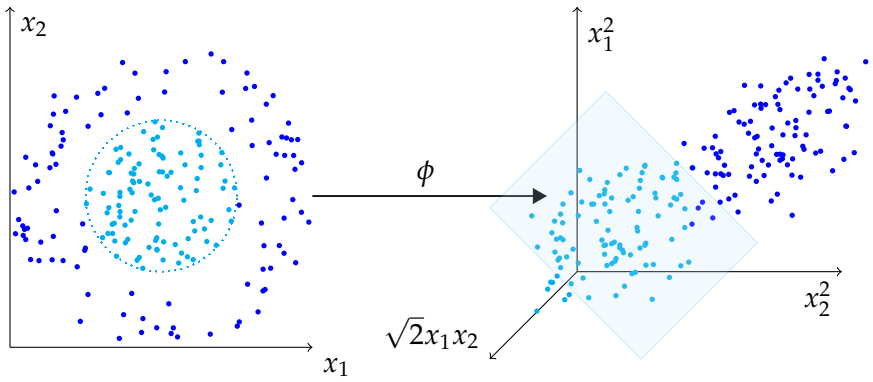

Figure 9: Example of polynomial feature mapping ϕ with degree two. Left: the data are not linearly separable in the input space. Right: the mapping to a higher dimensional representation allows the linear separation of the two classes.

feature map of degree two, we get a new data representation, through which the classes are linearly separable.

The prerequisite of having a feature map $\phi \in \mathcal{H}_k$, the reproducing kernel Hilbert space, guarantees a unique relation between the feature map and a semi-definite positive functional K, called kernel. The kernel measures the

pairwise distance between any sample in the $\mathcal{H}_k$ space [47]. This special property is also known as the *kernel trick*, as is widely used in several machine learning techniques [5, 84, 114]. This allows indeed to write any function equivalently through the kernel of the feature map. The former case is extremely convenient when the feature map is high dimensional, as it reduces to the computation of the similarity.

[110] Using the kernel trick, the formulation of the model in Equation 23 is equivalent to

$$f(\cdot) = BK(X, \cdot) + b, \tag{24}$$

with B the kernel coefficients to optimize, each weighting a sample of the dataset in X. In Table 4 some among the most popular kernels used in machine learning. The linear kernel corresponds to the identical feature map, the

Linear	$X_i \cdot X_j$
Affine	$(1 + X_i \cdot X_j)$
Polynomial	$(1 + X_i \cdot X_j)^m$
Gaussian	$\exp\left[-\frac{\|X_i - X_j\|_2^2}{\sigma^2}\right]$

Table 4: Examples of the most common kernels, pairwise similarity measures for the generic ith and j-th samples in the input space. For the linear, affine and polynomial mappings, the similarity is measured through the product, in the Gaussian kernel, it increases inversely to the norm of the distance.

affine K has an additional bias term, the polynomial kernel corresponds to the similarity matrix for the polynomial feature map, and the Gaussian kernel originates from the Gaussian function, with variance σ. This last feature map is infinite dimensional, but independently from the choice of the feature map, the desirable characteristic of kernel methods, given 24, is the bound to the model complexity given by the number of training examples [55].

3.4.2 *Non-linear methods*

A second class of more recent methods, which do not share the same mathematical properties of kernel machines, can be used to express non-linear function. Examples are *ensemble techniques* and *deep learning methods*. A large part of the community is investing its effort on this latter class of methods, as the non-linearity typical of the free parameters of these models is such that the solution strongly depends on initial conditions and convergence to an optimal solution (global optimum) cannot be always guaranteed [129], even if empirically observed [136]. A mathematical framework able to give general theoretical guarantees about these methods still do not exists, but there have been several attempts [102, 108]. Nonetheless, the approximation capability of these

models obtained through the minimization of a data fitting term seems the perfect learning machine (for more insight about the approximation capacity of non linear models, refer to the universal approximation theorem [30]).

3.4.3 *Learning Issues*

There are several aspects that may impact the results during the learning process, all related to the difficulty of dealing with real-world measures. A dataset may indeed be affected by noise due to the acquisition system, which could introduce random fluctuations in the measurements. Moreover N, the number of samples, could be not high enough to infer a model given the variables p under study. This is typically called over-parametrization or small N large p scenario. Lastly the presence of not relevant features to the learning task at hand may worsen the performance. For all these cases, the minimization of the approximation function could lead to an over-adaptation to the given set of points, resulting in poor predictive capacity. The phenomenon is called *overfitting*; in this case the model shows poor generalization capacities. The over-adaptation phenomenon, and poor predictive results obtained from these models become more dramatic as the models complexity increases.

We report an example of this phenomenon in Figure 10.

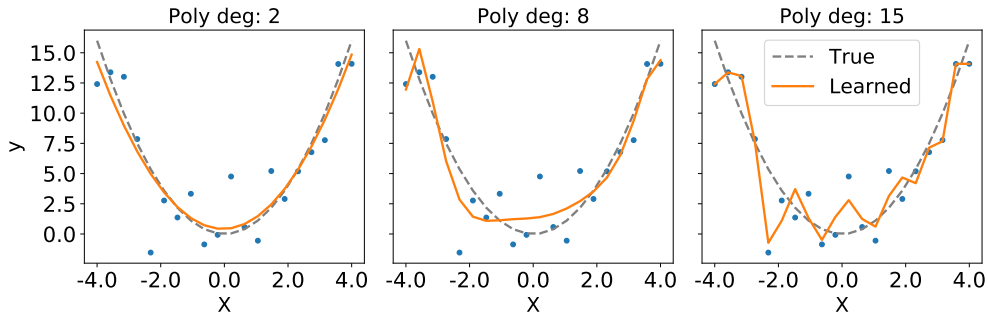

Figure 10: Example of overfitting for a regression task. The true input-output relation is $f(X) = X^2$, with Y affected by additive Gaussian noise. The noiseless f is the dashed gray curve. On the left: result of the fitting procedure using the least square loss, in the hypothesis of a polynomial of degree 2. In the middle: results from a polynomial of degree 8. On the right: fitting result from a polynomial of degree 15. The inferred model adapts better to the given data as the degree increases, but this is not a desirable behavior in terms of prediction for new unseen samples.

In the example we generate the true regression values Y using a parabolic function, affected by Gaussian noise. We fit a polynomial of degree 2, 8, and 15 for the three plots. The functions resulting from the learning procedure are reported in orange. As the model complexity increases we assists to the over-adaptation phenomenon, with evident irregularities of the learned model.

3.4.4 *Regularization*

As already seen in the approximation functionals in dictionary learning methods, regularization represents an effective strategy to (i) avoid over-adaptation of the model to the data, (ii) impose prior knowledge to our model.

Regularization techniques offer a way to mitigate the problem of overfitting by limiting the space of solutions to a smaller subset [40, 90], Chapter 5 [52]. The search of predictive models, in presence of regularization, can be formulated as the following minimization problem

$$\text{argmin}_f \left\{ \sum_{i=1}^{N} \mathcal{L}(Y_i; f, X_i) + \lambda \mathcal{R}(f) \right\}, \tag{25}$$

where we add to the *data-fitting term* a second term, called *regularization term* $\mathcal{R}$. Depending on the regularizer $\mathcal{R}$, we may favor smoother solutions and more regular functions. The parameter λ weights the importance of the regularization term.

Depending on the imposed constraint, the regularization term assumes different forms. In Table 5 we report the most popular choices for regularization terms.

L^1 norm	$\|f\|_1$
L^2 norm	$\|f\|_2^2$
Total Variation (TV)	$\|\nabla f\|_2$
Elastic Net (ENet)	$(1-\alpha)\|f\|_2^2 + \alpha\|f\|_1$

Table 5: Examples of typical $\mathcal{R}$ terms. The functional penalizes the use of a large number of coefficients for L^1, unstable models for L^2, strong differences between near coefficients in TV, and a mixture of L^1 and L^2 assumption for ENet.

With the sparsity assumption, meaning that not all the observed variables are informative for the predictive task, we suggest the choice of the L^1 term [51]. Another mild requirement on the solution is smoothness, obtained through L^2 regularization [90], which should limit abrupt changes in the prediction for small variations of the input data. To impose sparsity for correlated features, a good compromise is given by Elastic Net [34, 138]. The Elastic Net penalty merges indeed the contribution of L^1 and L^2 norms. If we aim at penalizing strong variations of the model for neighbor variables, then a Total Variation (TV) term, which measures the gradient norm [107], may be desirable.

We presented here a standard regularization framework, given by constraining the norm of the solution, but many other regularization techniques, developed to avoid overfitting, exist: early stopping [133], batch normalization [59], dropout [117] just to name a few.

3.4.5 *Model Selection and Model Assessment*

The choice of a model is a critic aspect, as we aim at finding a general law able to adapt to our learning data as well as to future data. Example of poor predictive models have already emerged, as reported in Figure 10. On one side, the use of rich models can lead to an over-adaptation of the model to the data, which is known as *overfitting*. On the other side, regularization can bound the problem complexity, but its contribution to the optimization problem must be tuned so not to have too strict models, which do not adapt to the data. This phenomenon is known as *underfitting*. The choice of the best hypothesis class and consequently of the best model is tricky. In this section we depict the main strategies used to avoid models with poor predictive performance, in favor of more general solutions.

We measure the model capacity to predict the outcome for future data through the *Generalization Error* (GE). Let $(X, Y)_{\text{ts}}$ denote the *test set*, or the input output variables extracted from the same distribution of the training data, $\mathcal{D}_{\text{tr}}$. The GE quantifies the expected prediction error or the discrepancy, among realizations of Y_{ts} and the prediction obtained through the model $\hat{f}$, as follows

$$\text{GE}_{\mathcal{D}_{\text{tr}}} = \mathbb{E}\left[\mathcal{L}(Y, \hat{f}(X))|\mathcal{D}_{\text{tr}}\right]. \tag{26}$$

Let us provide a direct interpretation of the generalization error through one of the most common cost functions, the least square loss. This loss penalizes the Euclidean distance between our model prediction and the true output. We make the hypothesis that our input-output relation is given by a determinist law f, as follows

$$Y = f(X) + \varepsilon,$$

with an additional random variable $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ which models the presence of noise. Let $\hat{f}$ be the relation inferred from a training dataset $\mathcal{D}_{\text{tr}}$. The generalization error translates to the expected loss evaluated on an independent dataset, given the model $\hat{f}$, which we consider fixed.

The generalization error for the square loss function can be written as follows

$$\text{GE}_{\mathcal{D}_{\text{tr}}} = \mathbb{E}\left[(Y_{\text{ts}} - \hat{f}(X_{\text{ts}}))^2\right] = \mathbb{E}\left[(f(X_{\text{ts}}) + \varepsilon - \hat{f}(X_{\text{ts}}))^2\right]. \tag{27}$$

In particular it decomposes in the sum of three terms which have direct interpretation.

$$\begin{aligned}\text{GE}_{\mathcal{D}_{\text{tr}}} &= \mathbb{E}\left[f^2(X_{\text{ts}}) + \varepsilon^2 + \hat{f}^2(X_{\text{ts}}) + 2\varepsilon f(X_{\text{ts}}) - 2\varepsilon\hat{f}(X_{\text{ts}}) - 2f(X_{\text{ts}})\hat{f}(X_{\text{ts}})\right] \\ &= \sigma^2 + f^2(X_{\text{ts}}) + \mathbb{E}\left[\hat{f}^2(X_{\text{ts}})\right] - 2f(X_{\text{ts}})\mathbb{E}\left[\hat{f}(X_{\text{ts}})\right]\end{aligned}$$

The generalization error can also be written as follows

$$\mathrm{GE}_{\mathcal{D}_{\mathrm{tr}}} = \underbrace{\sigma^2}_{\text{intrinsic error}} + \underbrace{\left(f(X_{\mathrm{ts}}) - \mathbb{E}\left[\hat{f}(X_{\mathrm{ts}})\right]\right)^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[\left(\hat{f}(X_{\mathrm{ts}}) - \mathbb{E}\left[\hat{f}(X_{\mathrm{ts}})\right]\right)^2\right]}_{\text{variance}} \tag{28}$$

The three terms quantify different contributions to the generalization error. The *intrinsic error* cannot be reduced, as related to the data acquisition process. The *bias* term gives a measure of the discrepancy of the model from the average prediction. In presence of underfitting we expect this term to dominate the generalization error. The *variance* quantifies the fluctuations of the predictor, we expect this term to be large in case of overfitting.

More in general the model complexity can be tuned thanks to the regularization parameter. As reported in Eq. 25, this requires the choice of a good trade-off between model complexity and data adaptation, and in this case depends on the value of the parameter λ. The curves in Figure 11 show the effect of the model complexity on the predictive performance of a learning method. The overfitting phenomenon is present as we increase the model complexity, right extremum in the spectrum of prediction performance, and can be measured as the gap between the training data and the one used to evaluate the goodness of the model, *validation set*.

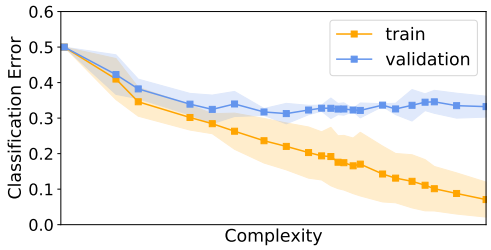

Figure 11: Example of overfitting phenomena, and the trade-off given by the obtained on multiple running of a learning method, as we increase its complexity. On the x-axis we report complexity, as the inverse of the regularization parameter, on the y-axis the classification error. The orange curve, corresponding to the training error is not a good estimate of the model performance. The blue curve does not show the same trend for increasing complexity.

The dots and areas in the plot refer to the mean value and standard deviation loss evaluated on the train data, in orange, and on some new validation data, in blue. The training error does not represent a good estimate of the validation error. As we increase model complexity the gap between the two curves becomes larger, leading to predictors with poor performance (overfitting). The complexity here is controlled by a multiplicative coefficient for the regularization term. When the effect of regularization is too strong, the model does not fit the training data, with poor predictive performance on both the training and validation set (under-fitting).

Through this example we notice more in detail that the choice of a good model requires us to fix several parameters. We can distinguish those in two families i) *hyper-parameters*, which are proper to the learning algorithm (e.g. scalars which weight the importance of the regularization terms, kernel's family, learning rates), ii) *model parameters*, which weight the variables measured during the data acquisition.

3.4.5.1 *Dataset splits*

In context where we are given an abundant dataset, the experimental protocol would be such to divide the dataset in three parts: the *training set*, the *validation set*, and the *test set*. The training set should be used to fit the model, the validation set serves to choose the best model among the ones proposed and to evaluate its robustness and stability, and a final test set is needed to assess the performance of the model on previously unseen data.

Nonetheless, in presence of a small amount of data, having a good estimate of the generalization error can be not trivial, due to random fluctuations. Other strategies can be preferable to assess predictive models and their performance with greater robustness. The evaluation protocol has a double goal: *model selection* and *model assessment*. Given a hypothesis class, the former task consists in the search, among the proposed models, of the one with the best performance. This is equivalent to identify the optimal hyper-parameters.

The model assessment serves to evaluate the robustness and stability of a model. We depict this process in Figure 12. The final output of the model assessment is the *expected generalization error*, in Eq. 29

$$\text{Exp GE} = \mathbb{E}\left[\mathcal{L}(Y, \hat{f}(X))\right] = \mathbb{E}\left[\text{GE}_{\mathcal{D}_{\text{tr}}}\right]. \tag{29}$$

The expected GE is the average error over different realizations of the experiment. We refer the reader to Section 7.12 [52] for further insight in the experimental design and the evaluation of generalization error or expected generalization error. In Figure 12 we show the procedure. We split learning and test data multiple times, one for each brown box. The learning set consists of both the training and validation sets; at this stage the choice of the best hyper-parameters and the best model, based on the performance on the validation set, takes place. The outcome of the brown box is a model $\hat{f}$ and the measure of the generalization performance over the test split. The average value of all these outcomes represents the expectation of the generalization error.

The split of learning and test sets, as well as for training and validation sets can be performed using several strategies. In Figure 13 we report the most common dataset splitting protocols: k-fold cross validation on the left, Monte-Carlo cross validation in the middle, and bootstrap.

3.4.5.2 *k-fold Cross Validation*

The procedure consists in splitting the dataset in k non intersecting subsets. We used $k-1$ subsets to learn a model and we leave out one split as an evaluation

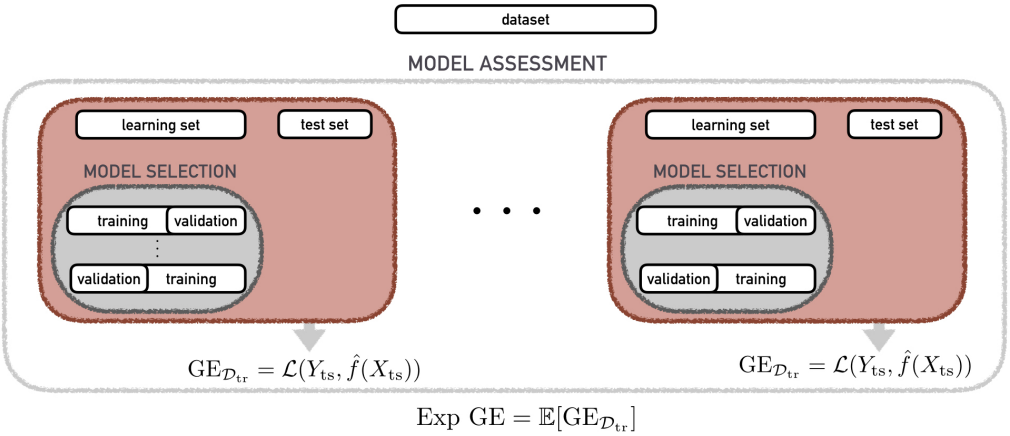

Figure 12: Model assessment procedure for the small N scenario. Given a dataset, we split the learning and test sets multiple times. At each repetition, represented through the red box, we further split the learning set several times, to select the best model, including the set of hyper-parameters which regulates its complexity, grey box. We retrain on the entire learning set and we estimate the performance on the test set. We measure the $\mathrm{GE}_{\mathcal{D}_{\mathrm{tr}}}$, even if we are in the small sample size scenario. We compute the overall performance Exp GE.

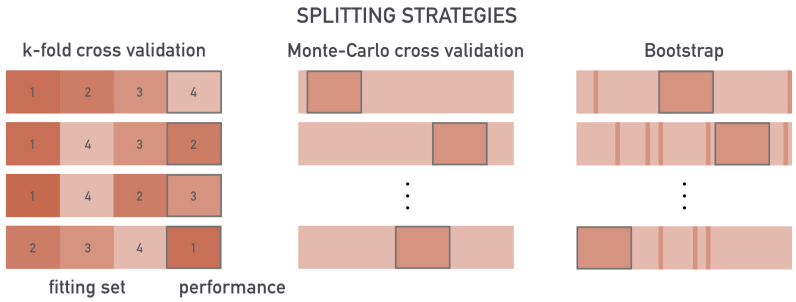

Figure 13: Dataset splitting strategies used in model assessment and model selection for a 4-fold cross validation. On the left, k-fold cross validation is based on non intersecting k splits of the dataset. In the middle, Monte-Carlo cross validation separates the evaluation and test set, without contaminations. On the right, bootstrap, where evaluation and learning set are proportion of the entire dataset, sampled with repetition.

set. We repeat this procedure for all the different k folds. The prediction error is estimated as the average of the performance obtained on the left out split. In Figure 13, we report the example for $k = 4$. The case $k = N$ is also known as *leave-one-out* cross validation.

3.4.5.3 *Monte-Carlo Cross Validation*

We fix ν, the proportion of evaluation data and we randomly split the data in two sets, the evaluation set, consisting of $n\nu$ samples, and $n(1 - 1/\nu)$ samples as learning set. This procedure can be repeated multiple times, differently from

the k-fold cross validation scheme. At each repetition the performance of the model learned are evaluated on the evaluation set.

3.4.5.4 *Bootstrap*

We fix a proportion on evaluation samples and we sample with replacement the learning and the evaluation set from the entire dataset. This may lead to scenarios with some samples contained in both the learning and evaluation sets. For this reason this approach is typically used in the unsupervised setting.

3.4.5.5 *Hyper-parameters search*

The definition of the hyper-parameters search space can be performed through a *grid search* or *random search*. In grid-search, given the hyper-parameters of the model, we define for each an array of possible values. We then generate the tuples of all their combinations, and we ran the k-fold protocol for each tuple. The best tuple $(\lambda_1, \ldots, \lambda_J)^*$ is selected as the one which obtains the best average score on the validation set. In random search, we define a distribution for each hyper-parameter. For a fixed amount of times, we draw the hyper-parameters tuple and we compute the performance of the model with the k-fold cross validation scheme. Again, we select the tuple corresponding to the best performance $(\lambda_1, \ldots, \lambda_J)^*$. While grid-search is an exhaustive search criterion for model selection, random search is more flexible as it fixes the total number of tuples of parameters. For this reason random search is more suitable to large datasets, to models with large number of hyper-parameters, and in case of limited computational resources [13].

3.4.6 *Classification Methods*

As specified at the beginning of this section, there are two main tasks of supervised learning, classification and regression, but we focus on the former, as we will deal only with classification problems throughout this work. We introduce linear methods e.g. logistic regression, support vector machines, which can be extended to non linear models through features maps. Then we consider ensemble methods, as random forests and gradient boosting. We report non-linear classification methods as neural networks.

3.4.6.1 *Logistic Regression*

Logistic regression (LR) is a common classification method which models the posterior probability through a linear function [52]. We introduce the binary case only, even if the generalization to the multi-class scenario is straightfor-

ward. Let the labels for the two classes be coded as $\{\pm 1\}$. For the generic sample i, its probability to belong to each class is modeled as

$$\Pr(Y = +1|X = X_i) = \frac{\exp[\beta_0 + X_i\beta]}{1 + \exp[\beta_0 + X_i\beta]} \tag{30}$$

$$\Pr(Y = -1|X = X_i) = \frac{1}{1 + \exp[\beta_0 + X_i\beta]} \tag{31}$$

Given N samples, the likelihood for N independent observations corresponds to

$$\text{likelihood}(X, Y) = \Pi_{i=1}^{N} \Pr(Y = Y_i|X = X_i), \tag{32}$$

or equivalently

$$\text{likelihood}(X, Y) = \Pi_{i=1}^{N} \left(\Pr(Y = +1|X = X_i)\right)^{\frac{1+Y_i}{2}} \left(\Pr(Y = -1|X = X_i)\right)^{\frac{1-Y_i}{2}}.$$

As the search of the optimal parameters from the maximization of the likelihood gives the same result of logarithm of the objective, we consider

$$\log \text{likelihood}(X, Y) = \sum_{i=1}^{N} \left(\frac{1 + Y_i}{2} \log(\Pr(Y = +1|X = X_i)) + \frac{1 - Y_i}{2} \log(\Pr(Y = -1|X = X_i)) \right)$$

By substituting the definitions 30 in the Formula, we obtain

$$\log \text{likelihood}(X, Y) = \sum_{i=1}^{N} \left(\frac{1 + Y_i}{2}(X_i\beta + \beta_0) - \log(1 + \exp[\beta_0 + X_i\beta]) \right). \tag{33}$$

The solution to this problem is typically reached iteratively through Newton-Raphson algorithm. As the method returns the optimal tuple (β^*, β_0^*), the label for a new test point X^{test} is assigned by considering for which of the two probabilities $(\Pr(Y = +1|X = X^{\text{test}}), \Pr(Y = -1|X = X^{\text{test}}))$ we get the maximum value.

3.4.6.2 *Support Vector Machine*

Support Vector Machine (SVM) is a popular regularized binary classification method [29], which has been lately extended to the multi-class case [101]. Given a binary classification problem SVM aims at finding, among all the possible solutions, the linear function which separates the two classes at best. The optimality and uniqueness of the SVM solution is given by the *margin maximization*. The idea behind SVM is geometrically intuitive, as reported in Figure 14. Orange and blue dots represent samples from the two classes. The hyperplane which optimally splits the space is the one which maximes the margin, corresponding to the width of the yellow area.

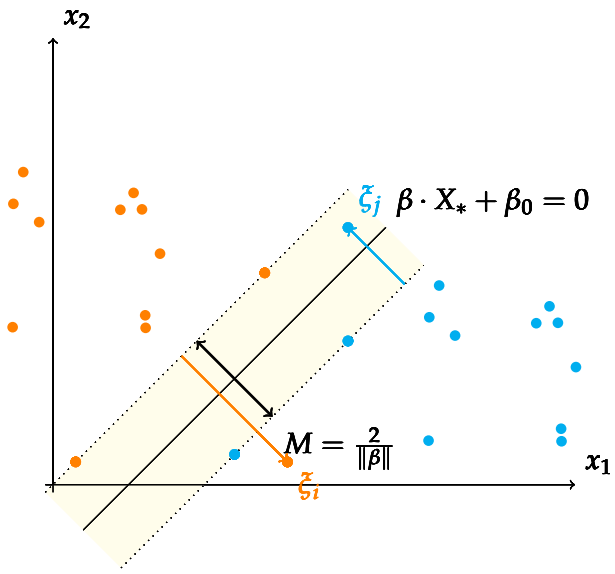

Figure 14: Linear separating hyperplane in a two dimensional feature space. The task reduces to find the best linear separating function, or the one that maximizes the distance between the two classes. The method takes into account the presence of errors through the variables ξ_*, which weight the misclassified examples through the hinge loss function in Tab. 3.

Given the linear function $\beta \cdot X_* + \beta_0 = 0$, as a candidate separating hyperplane, the equations relative to the margin (dotted lines) can be written as

$$\beta \cdot X_* + \beta_0 = +1, \text{ positive class}$$
$$\beta \cdot X_* + \beta_0 = -1, \text{ negative class.}$$

The distance between the two hyperplanes is equivalent to the quantity $2/\|\beta\|$, also known as the *margin*. SVM can also take into account the presence of outliers. Referring to Figure 14, the outliers are here the two dots which fall in the wrong side of the margin. The presence of outliers is penalized through the hinge loss which quantifies the distance of these points from their proper subspace. For each sample i, we define the correspondent hinge loss evaluated at that point as slack variable ξ_i, with

$$\xi_i = \max(1 - Y_i(X_i\beta + \beta_0), 0). \tag{34}$$

The SVM optimization problem in the non-separable scenario corresponds to the minimization of the following objective function

$$\text{argmin}_{\beta,\beta_0,\{\xi_1,\cdots,\xi_N\}} \left\{ \frac{\|\beta\|}{2} + C \sum_{i=1}^{N} \xi_i \right\}, \tag{35}$$

$$\text{subject to: } Y_i(X_i\beta + \beta_0) \geq 1 - \xi_i, \tag{36}$$

$$\xi_i \geq 0, \forall i \in \{1, \ldots, N\} \tag{37}$$

where the first term is also a regularization term, the second term corresponds to the approximation term, and C regulates the trade-off and is equivalent

to $1/\lambda$ of the regularized methods seen in Formula 25. The presence of constraints over the ξ variables leads to a quadratic problem in the Lagrangian variables, in its dual formulation (see Section 12.2.1 of Hastie et al. [52]). Such problem is convex, thus we can always obtain a solution denoted as (β^*, β_0^*). Given a new test sample, the label is assigned by taking into account the output of the sign function

$$y^{\text{test}} = \text{sign}(X^{\text{test}}\beta^* + \beta_0^*). \tag{38}$$

3.4.6.3 *Decision Trees and Ensemble Methods*

Decision trees are non-linear methods widely used in classification. They consist in finding charts of decision which, given a sample, guide through its structure until the class is assigned. Decision trees operate a partition of the input space in subset of rectangles. For sake of clarity we report an example in a two dimensional feature space in Figure 15. In the example at the root we

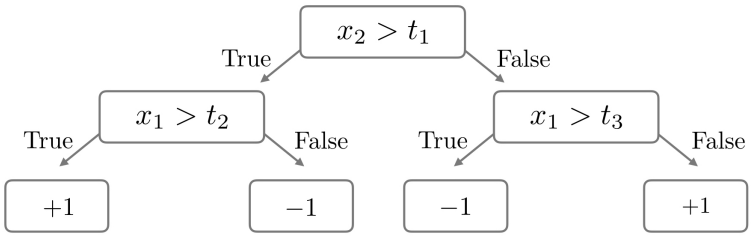

Figure 15: Example of a decision tree in a two dimensional feature space. The cuts are operated through three scalar quantities t_1, t_2, t_3. Given a new sample, we follow the chart to find the classification output.

have the variable x_2. The choice of the tuple $\theta = (k, t_m)$, with k denoting the k-th feature and t_m the cutting at node m is the result of an optimization procedure at each step. Let $\mathcal{S}_{m,1}$ and $\mathcal{S}_{m,2}$ denote the two general subsets obtained at the m-th node. The search for the optimal parameters is performed through a greedy procedure.

$$\mathcal{S}_{m,1}(\theta) = (X_i, Y_i) : (X_i)_j < t_m$$
$$\mathcal{S}_{m,2}(\theta) = (X_i, Y_i) : (X_i)_j > t_m$$

Let $n_1 = \#\mathcal{S}_{m,1}(\theta)$ be the cardinality for the set 1 at node m, and $n_2 = \#\mathcal{S}_{m,2}(\theta)$, the cardinality for the set 2 at node m, and n_m be the sum of the two. The optimization problem translates to

$$\theta^* = \text{argmin}_\theta \left\{ \frac{n_1}{n_m} H(\mathcal{S}_{m,1}(\theta)) + \frac{n_2}{n_m} H(\mathcal{S}_{m,2}(\theta)) \right\} \tag{39}$$

where H denotes an *impurity function*, which measures the discriminative capacity of the cut. Typical impurity functions are Gini criterion and the information gain criterion. The former quantifies the goodness of a split by evaluating

the probability of labeling a random sample obtained at the split m as the one of the correct class

$$H(\mathcal{S}_{m,1}(\theta)) = p_1(1-p_1) + p_{-1}(1-p_{-1}), \tag{40}$$

with $p_1 = \#\{(X_i, Y_i) \in \mathcal{S}_{m,1}(\theta) | Y_i = 1\}/n_1$ and $p_{-1} = \#\{(X_i, Y_i) \in \mathcal{S}_{m,1}(\theta) | Y_i = -1\}/n_1$. The information gain criterion consists in the maximization of the cross entropy

$$H(\mathcal{S}_{m,1}(\theta)) = p_1 \log_2(p_1) + p_{-1} \log_2(p_{-1}), \tag{41}$$

with p_1 and p_{-1} defined as before. The non linearity of the decision tree classifiers emerges already from the example in Figure 15. The partition of the input space for this case is pictorially reported in Figure 16. Decision trees suffer of

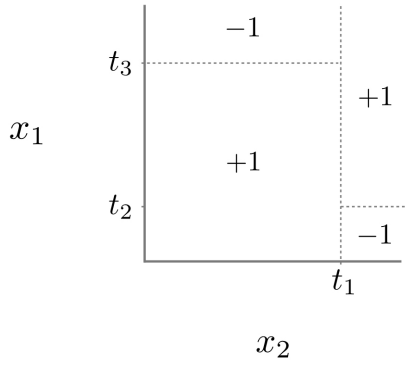

Figure 16: Partition of the input space for the chart of the previous example. The values t_1, t_2 and t_3 determine consecutive cuts in the space and lead to the optimal division of the space

some weaknesses such as instability and lack of robustness in the presence of noisy measures, which derive from the hierarchical structure of the prediction. Different techniques have been proposed to solve these issues, which relies on predictions from multiple estimators and go under the name of *ensemble methods*.

RANDOM FOREST First introduced by Breiman [19] Random Forest is an aggregative strategy based on bootstrap. Through this method B decision tree models are trained in parallel using a subset of the entire dataset, sampled through a bootstrapping strategy. The random forest output for each data point consists in B predictions. The classification label is assigned through majority voting. For further details about the decrease of the model variance given by this aggregating procedure is quantified in Section 8.7 (Bagging [52].)

GRADIENT BOOSTING Gradient Boosting is an ensemble technique based on the sequential training of a set B of predictors [43]. The procedure is such that, for each new model, the data are weighted depending on the misclassification committed at the previous step, so to force the algorithm to learn the input-output relation previously underrated. The iterative improvement given by GB can be seen as a functional gradient descent [81].

3.4.6.4 *Neural Networks and Deep Learning*

Recently, the scientific community assisted to an explosion of automated classification tools, in image classification, time series analysis, reinforcement learning, and generative models which leverage on deep learning [70, 87, 95]. This terms collects a plethora of complex architectures for which mathematical properties e.g. stability, robustness, and convergence to general solutions are difficult to prove. Entire books have been written for each section of this chapter, including deep learning [24, 46], so here we only give a short overview into the main concepts. We start by considering the ancestor of deep learning architectures, which are shallow neural networks, up to more complex and highly overparametrized methods.

In the last part we introduce Convolutional Neural Networks, as those represent a good compromise between supervised learning and data representation. Indeed deep learning methods first serve the supervised task, solving classification or regression problems, but to their flexibility, they act as memory units with the side effect of returning useful, sometimes compressed data representations. In this regard, they can also be used as tools for feature extraction, as the extracted features may be general enough to serve other future learning tasks [12].

ROSENBLATT'S PERCEPTRON represents the fundamental unit of a complex neural network. The learning unit consists of a non-linear function σ (e.g. a sigmoid, a rectifier, or a step function) which takes as argument a linear combination of the input data. Given the collection of N training samples $\{X_i, Y_i\}_{i=1}^N \in \mathbb{R}^p \times \mathbb{R}$ the perceptron model aims at approximating the input output relation as

$$f(X) = \sigma(w \cdot X), \tag{42}$$

with σ a non linear function.

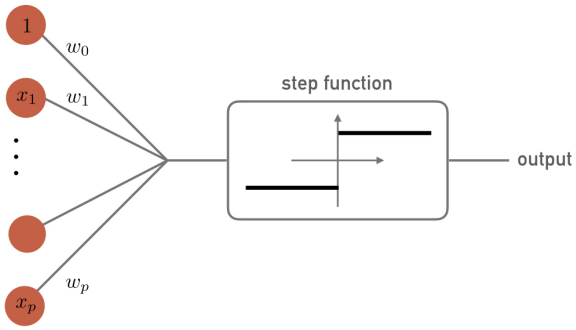

Figure 17: Single neuron architecture. The non linear function corresponds to the step function, its argument is a weighted combination of the features from the input data.

The search of optimal parameters takes place through the minimization of a loss function, which is a user choice. People refer to perceptron as a single

unit, as the later extension of network methods relies on multiple use of single units or/and on concatenation of single units.

MULTILAYER PERCEPTRON With neural networks the community typically refers to the stack of multiple layers of parallel single units, which cooperate to the same learning task. The first layer linearly combines the input features, and the following layers take as input the output of the previous layer. Shallow networks have a single hidden layer. The operation performed for each hidden node is equivalent to the one described for the perceptron (see Eq. 42). The novelty of this approach is due to the presence of hidden layers. The formula for a single hidden layer corresponds to the following

$$f_i(X) = \rho\left(\sum_{k=1}^{p} w_{ik} X_k\right), \; i \in \{1, \dots, h\} \tag{43}$$

$$Y_j = \sum_{i=1}^{h} \beta_{ji} f_i(X), \; j \in \{1, \dots, o\}, \tag{44}$$

while, in the multilayer scenario, we have the multiple composition of non linearities, alternated to linear maps, as in Figure 18.

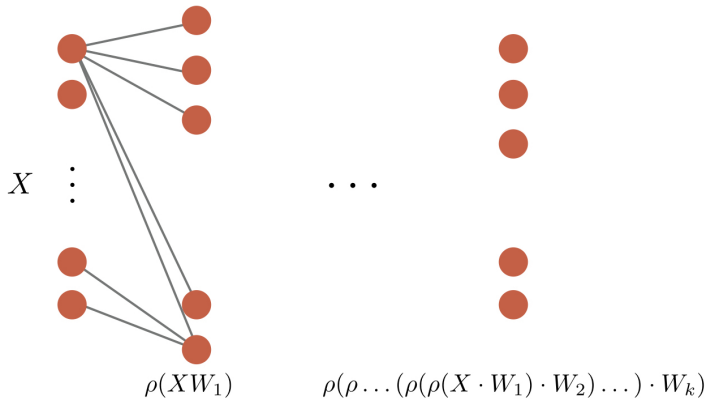

Figure 18: Multilayer perceptron architecture. At the first layer the input features are linearly combined to form a single output, and go as input to a node of the hidden layer. This procedure is repeated multiple times, depending on the depth of the architecture. The last layer collects the feature representations obtained through the network and is key for the prediction task.

The set of weights for the model are found through the minimization of a loss function, which is again a user's choice. Given a loss $\mathcal{L}$ the optimization procedure relies on *back propagation*.

At each iteration the update of the solution is performed in two steps: the *forward step* which consists in fixing the weights to compute predictions and the *backward step* where the updates for the outer layers are applied and propagated to the inner layers weights. The interesting property of this procedure is given by the selective nature of the updates. Through this sort of cascade of updates, each hidden unit passes information and receives inputs only

from nodes which share a connection with it. The computational advantage of this technique is given by the possibility of implementing the neural network on parallel architectures, specifically designed for scientific computation, as GPUs. This represents one of the main reasons for their widespread use nowadays.

CONVOLUTIONAL NEURAL NETWORKS An emblematic case of learning a representation by imposing a learning task is provided by transfer learning, typical in Convolutional Neural Networks (CNNs).

The CNNs breakthrough has been possible thanks to the technological advancements of hardware, e.g. GPUs and TPUs, which gave a great acceleration to the computational speed of these overparametrized models.

CNNs rely on the composition of a high number of hidden units, where the mathematical operations across layers are extremely general and rely on convolutions. Given an image or a time series, at each layer the weights to optimize are respectively two dimensional patches, or one dimensional arrays convolved with the output of the previous layer. In complex architectures, several operations across layers are performed for dimensionality reduction, e.g. max-pooling or average-pooling [6]. One of the first deep neural network used for the ImageNet data classification has been AlexNet, in Figure 19.

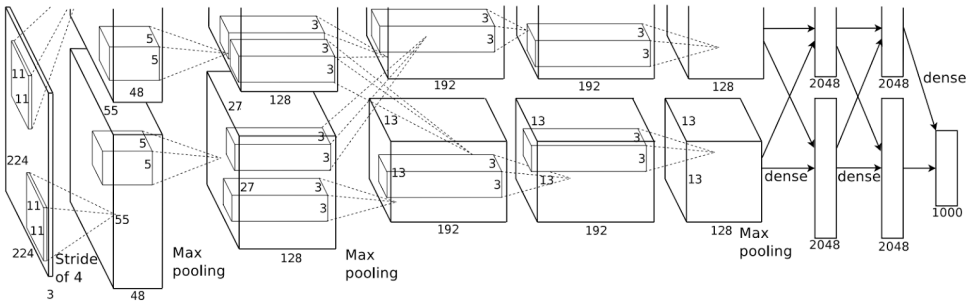

Figure 19: As the computational capacity grows, the possibility of building complex learning architecture explodes, leading to multilayer models. Here it is shown AlexNet, a convolutional neural network which competed and won the ImageNet Large Scale Visual Recognition Challenge in 2012 [70]

Given the models complexity, providing a theoretical framework for these architectures becomes unfeasible. The formulation of a theoretical framework for deep learning has been the focus of a consistent part of the machine learning community, as these models show good predictive performance despite being heavily overparametrized [136].

The learned filters in CNNs can compete with the engineered ones. An example emerges already from the work of Krizhevsky, proposed in Figure 20. Here 96 filters, RGB squares of edge 11 pixels, show similarity with handcrafted filters as Gabor or Wavelets in two dimensions. The filters in the first layers of the architecture have extremely simple shapes, which are composed at the higher layers of the representation to generate more complex patterns. These features can serve as a tool for feature extraction, and in the case of

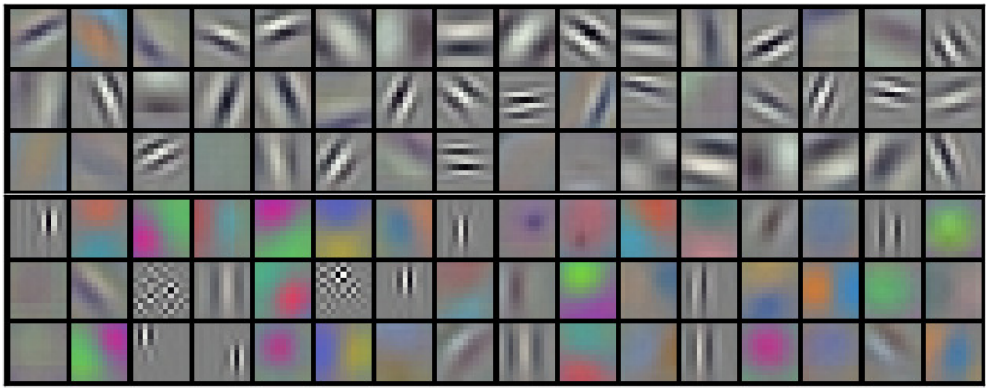

Figure 20: Image from the work of Krizhevsky et al. [70]. The 96 convolutional kernels with dimension $11 \times 11 \times 3$ at the first hidden layer of AlexNet. The filters shape reminds famous Gabor and wavelet filters in two dimensions.

small number of training example, it is common to train shallow models on top of data-driven features extractor as CNNs layers.

3.4.7 *Evaluating Models Performance*

The choice of a metrics is fundamental to evaluate the predictive performance of a model. This is typically realized by comparing the predicted label $\hat{Y}$ to the true one Y, for all the samples in the evaluation set. In binary classification this leads to four possible outcomes: True Positive (TP), predicted and true labels are both positive, True Negative (TN), predicted and true labels are both negative, False Positive (FP), for negative samples with positive prediction, and False Negative (FN), for positive samples with negative prediction.

In Table 6 we report some of the most popular classification scores. The metrics proposed here assume continuous values between $[0, 1]$. Accuracy is not an optimal choice for unbalanced datasets. In this case indeed the random prediction score does not corresponds to 0.5, but must be fixed depending on the classes unbalance. The other scores are sensitive to the classes unbalance and allow a further characterization of the model performance.

3.5 Learning Methods for Clustering

In absence of target labels, clustering methods aim at grouping sample points based solely on their distances. As such, these fall in the class of the unsupervised learning methods. The choice of a proper distance is a critic aspect and it is based on prior assumptions [63]. Many algorithms have been designed for this task, from the more fundamental *k-means* or *centroid based clustering* to more complex methods as *hierarchical clustering* and *spectral clustering*. All of

Classification Metrics		
Accuracy	ACC	$\frac{1}{N} \cdot (\mathrm{TP} + \mathrm{TN})$
Balanced Accuracy	BALACC	$\frac{1}{2}\left[\frac{\mathrm{TP}}{\mathrm{TP+FN}} + \frac{\mathrm{TN}}{\mathrm{TN+FP}}\right]$
Precision	P	$\frac{\mathrm{TP}}{\mathrm{TP+FP}}$
Recall	R	$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$
False Positive Rate	FPR	$\frac{\mathrm{FP}}{\mathrm{TN+FP}}$
False Negative Rate	FNR	$\frac{\mathrm{FN}}{\mathrm{TP+FN}}$
True Negative Rate	TNR	$\frac{\mathrm{TN}}{\mathrm{FP+TN}}$
F1 score	F1	$2\frac{\mathrm{P \cdot R}}{\mathrm{P+R}}$

Table 6: List of the most common score metrics for classification. The acronyms TP, TN, FP, and FN denote respectively true positive, true negative, false positive and false negative samples based on the model prediction. Accuracy is not an optimal choice in presence of unbalanced classification problems.

them rely on the evaluation of a dissimilarity matrix D which quantifies the pairwise distance among samples.

$$D = \begin{bmatrix} 0 & & d_{1N} \\ & \ddots & \\ d_{1N} & & 0 \end{bmatrix} \tag{45}$$

An exhaustive review of clustering methods is presented in Jain et al [63]. There, clustering approaches are categorized as *hierarchical* and *partitional*. In the former case a nested tree of dependencies *dendrogram* is built. In the latter case the partition is operated at once.

3.5.1 *Hierarchical Clustering*

The similarity measures are computed pairwise, then they are ranked is ascending order. The method does not require a number of clusters but a measure of dissimilarity across clusters. The algorithmic approaches can be divisive or agglomerative: in the former at the first iteration each sample constitutes a cluster (top-down), in the latter at the first iteration all the samples belong to the same cluster and the split is operated recursively (bottom-up). The choices to be made in the hierarchical approach concern: i) the metrics used to measure similarity across samples; ii) the linkage relation across different clusters, which determines the shape of the clusters and their partition.

Different linkage types are enumerated in the following

(i) single, $L(C_1, C_2) = \min_{x_i \in C_1, x_j \in C_2} \{\mathrm{distance}(x_i, x_j)\}$,

(ii) complete, $L(C_1, C_2) = \max_{x_i \in C_1, x_j \in C_2} \{\text{distance}(x_i, x_j)\}$,

(iii) average, $L(C_1, C_2) = \frac{1}{\#C_1 + \#C_2} \sum_{x_i \in C_1} \sum_{x_j \in C_2} \text{distance}(x_i, x_j)$,

(iv) ward, $L(C_1, C_2) = \frac{\#C_1 \#C_2}{\#C_1 + \#C_2} \|\mu_{C_1} - \mu_{C_2}\|$,

where C_1 and C_2 denote the two clusters, $\#C_1$ and $\#C_2$ their cardinality, μ_1 and μ_2 their centroids. In (i), the distance between two clusters is equivalent to the minimum distance between all the pairwise distances of the point belonging to the two different clusters. This choice is known to suffer from the *chaining phenomenon*, or the generation of clusters which rise from a series of concatenated close observations. The diameter of the clusters may be very large [52]. In opposition to this approach there is the complete linkage (ii), here the distance between two clusters is equivalent to the maximum distances between all the pairwise distances of the point belonging to the two different clusters. A previous work showed the capacity of constructing more compact clusters through this latter method. Both [52] and [63] agree on the higher compactness and stability of linkage type (ii) over (i). The approach (iii) is based on average and represents a compromise between (i) and (ii). The linkage type (iv) consists in the minimization of the variance within elements from the same cluster.

3.5.2 *Partitional Clustering*

When hierarchical clustering is unfeasible due to memory constraints, partitional clustering represents a valid alternative. A typical and most basic example of partitional clustering is the k-means algorithm (Voronoi iterations). Here, given k, number of clusters, a new sample is assigned to the cluster with the nearest mean. Given a set N of observations and a number k of clusters, the related optimization problem is defined as

$$\text{argmin}_{\{C_1, \dots, C_k\}} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|_2^2 \tag{46}$$

Given an initialization, which consists in a set of $\{\mu_1^{(1)}, \dots, \mu_k^{(1)}\}$ centroids, the algorithm consists of two steps, the *assignment step*, which quantifies the distance between the generic i-th sample and the centroids and assigns it to the cluster with the smallest distance, and the *update step*, where the new centroids are computed based on the previous assignments.

3.5.3 *Evaluating Goodness of Clustering*

There are several metrics which evaluate the goodness of clustering methods, but some of them make use of ground-truth labeled samples. Some examples are *adjusted rank index*, *mutual information score*, *V-measure*, and *homogeneity*. As we typically lack information about the true labels, we report a metric

which does not rely on those, but rather represents a measure of geometrical properties from the resulting clusters.

3.5.3.1 *Silhouette Score*

Assuming the existence of K clusters and a distance metric d, we define the measure a for each point in the generic cluster i

$$a(i) = \frac{1}{\#C_i - 1} \sum_{k \in C_i,\ i \neq k} d(X_i, X_k) \tag{47}$$

as the mean pairwise distance among points belonging to the same cluster.

A second quantity defined for each i point is given by considering the distance between i and the mean distance with all the points belonging to another cluster k.

$$b_k(i) = \frac{1}{\#C_k} \sum_{s \in C_k} d(X_i, X_k). \tag{48}$$

The minimum distance of a generic point i from cluster C_i to another cluster is defined as

$$b(i) = \min_{o \neq i} b_o(i). \tag{49}$$

The silhouette value of each point in the dataset is defined as

$$s(i) = \begin{cases} \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, \text{ if } \#C_i > 1, \\ 0, \text{ otherwise.} \end{cases} \tag{50}$$

The silhouette score assumes value in the interval $[-1, 1]$, with 1 for perfect clustering algorithm.

PART III

Investigation and Main Experiments

4

Dataset Description

4.1 Questions and Motivations

In the introduction we briefly described focal epilepsy and the problem of EZ localization through invasive recordings. As reported in Chapter 2, we pointed out the existence of a plethora of works aimed at assessing the contributions of several electrophysiological waveforms and neurophysiological rhythms as biomarkers of the pathology. A large number of methods consider only a single or a small subset of candidate patterns of epileptogenicity. Even though this represent a reasonable approach for the definition of new biomarkers, it does not allow to have a general understanding of their relative importance in the localization of the EZ, neither to assess their collective predictive power.

On the other hand, the manual tagging of waveforms can potentially be a biased procedure, as performed through visual inspection and rarely guided by automatic tools. The discrimination between epileptogenic and physiological activity requires a high-level expertise, and this in general represents a limit for the reproducibility of these studies.

4.2 SEEG Data from Focal Epileptic Population

Our population consists of 60 subjects, all but one suffering of drug resistant focal epilepsy. All the subjects underwent the presurgical evaluation through SEEG.

The data were acquired at the Centre of Sleep Medicine, Centre for Epilepsy Surgery, Ospedale Ca' Granda Niguarda, Milano, Italy[1]. All the subjects signed the written consent to further analysis for scientific purposes. We dispose of invasive electrophysiological recordings relative to the *wakeful resting interictal stage* for all the 60 subjects.

Neurologists registered local field potential with common reference in white matter, using platinum-iridium, multi-lead electrodes. The number of contacts for each electrodes varies from 8 to 15, each is 2 mm long, 0.8 mm of thickness

1 Last access: May 10th, 2020 `https://esrs.eu/laboratory/department-of-neuroscience/`

and have distance of 1.5 mm from its neighbor contacts (DIXI medical, Besancon, France). The acquisition system is a 192-channels SEEG amplifier system (NIHON-KOHDEN NEUROFAX-110).

The acquisition protocol is such that the patient lays down, in a state of wakefulness with closed eyes, and in absence of external stimuli. This is also known as resting state. These registrations corresponds to interictal activity, the furthest from the epileptic seizure. We have been provided with recordings acquired simultaneously on multiple sites. The length of the registrations varies among patients, but it is comprised between the 10 and 15 minutes, at a sampling frequency of 1 kHz.

The neurophysiologists evaluated each recording during the presurgical phase, and annotated each monopolar contact. Contacts positioned in the *epileptogenic zone are labelled as* $+1$, while contacts in the *non epileptogenic zones are labelled as* -1. Throughout the work we refer to this evaluation as *pre-surgical assessment*.

The passage to the bipolar montage can be considered a spatial gradient of the signal. It moreover requires to adjust the binary annotations of the monopolar montage. The conversion is operated in such a way that, if at least one between two neighbor recordings (C_k, C_{k+1}) was tagged as recorded from an epileptogenic zone, we label it as epileptogenic, or pathological. This translates to the following

$$Y^{(p)}_{C_k - C_{k+1}} = \begin{cases} -1, \text{ if } Y^{(p)}_{C_k} = Y^{(p)}_{C_{k+1}} = -1 \\ +1, \text{ otherwise.} \end{cases}$$

The index p defines the generic p patient.

No annotation has been provided regarding the presence of pathological biomarkers or epileptic waveforms in the recordings.

For a subset of 40 patients we dispose of geometrical coordinates of the montage. This information has been extracted through FreeSurfer [42]. The exact positions of the implanted electrodes is measured using a fused MRI-pre with CT-post using affine rigid-body co-registration. After the co-registration phase, the algorithm automatically segments each contact contained in the electrodes by searching the center of mass for each contact [22]. The results of the segmentation algorithm are, for each contact, the *assigned anatomical region*, based on the Destrieux atlas[2], the *geometrical contacts positions* in cartesian coordinates [mm], the *Partial Tissue Density* (PTD) [83] and the *Grey-Matter Proximity Index* (GMPI) [89], which are measures of proximity to grey matter. As this information is not homogeneous across the entire population it will be reported in the following through a binary label.

For sake of clarity, the pharmacological treatment administered to each subject is provided. The treatment variability across the entire population is high. The acronyms in the Tables correspond to the following AntiEpileptic Drugs (AEDs): CBZ *carbamazepine*, CLB *clobazam*, CLZ *clonazepam*, DT *dintoina*, FB *fenobarbital*, FN *phenytoin*, LBT *missing*, LM *lamotrigine*, LS *lacosamide*, LR *lo-*

2 Last access: May 10th, 2020 `https://surfer.nmr.mgh.harvard.edu/fswiki/DestrieuxAtlasChanges`

razepam, LV *levetiracetam*, NTZ *nitrazepam*, OXC *oxcarbamazepine*, PRM *primidone*, RP *rufinamide*, SR *sertraline*, TP *topiramate*, TPM *topamax*, VP *valproic* and ZN *zonisamide*. The dosage for each AED is given in milligrams [mg]. Surgical and post-surgical information is also available. This information has been provided during the project, at the beginning of 2019. They are indicated in the Tables as (*). They concern the type of surgery for each subject, Ablation or Thermocoagulation, reported in the entry (A/T). In case where the surgical intervention has not been performed, we use the None flag. The region removed through ablation, or subjected to thermocoagulation are provided, together with the result of the post-surgical classification (Engel classification).

I use the notation - for missing entries.

In Tables 8, 9, 10, 11, 12, 13 we report the main characteristic of the dataset.

Table 7: A short guide to the Tables columns.

Acronym	**Full name**
#C	number of bipolar recordings
#PC	number of pathological bipolar recordings
P	availability of the contacts position
A/T	ablation/thermocoagulation
Region	ablated/thermocoagulated brain areas
Engel	Engel class
AED [mg]	Pharmacological treatment at the time of the SEEG

Table 8: First batch of patients. The columns report respectively: Subject ID subject identification number, #C the total number of recordings in the bipolar montage, #PC the number of pathological channels, P a binary label which defines the presence (Y) or absence (N) of the spatial information, A/T surgical intervention type, which differs in ablation or termocoagulation, Region the brain area removed through surgery, Engel post-surgical classification, AED, administered pharmacological treatment during the SEEG acquisitions.

Subject ID	#C	#PC	P	A/T*	Region*	Engel*	AED* [mg]
1	91	13	Y	A	right mesial frontal	IB	CBZ 600, LTV 1k
2	129	40	Y	A	right temporal insular	IA	CBZ 1.2k, PRM 750, CLZ 10
3	147	11	Y	A	right temporal	-	CBZ 1k, DT 450,LS 300, CLZ 10
4	125	51	N	A	left temporal parietal	IA	FB 100, TP 100, LV 3k
5	158	27	N	A	left temporal insular	IA	OXC 600, LS 400
6	114	37	N	A	right temporal insular	IIIA	LBT 1k, LS 350, SR 50, LR 1
7	156	51	N	-	-	-	-
8	127	33	Y	T	left temporal orbital	IA	CBZ 1.2k, LV 3.5 k, TPM 200
9	132	23	Y	None	None	None	OXC 600, LS 400
10	137	36	N	-	-	-	-

Table 9: Second set of patients. The columns are the same as in Table 8

Subject ID	#C	#PC	P	A/T*	Region*	Engel*	AED* [mg]
11	149	11	Y	-	-	-	-
12	119	29	Y	A	left temporal	IA	TP 200, CBZ 900
13	157	18	N	A	right anterior temporal	IIA	CBZ 1.2k
14	152	36	N	-	-	-	-
15	140	52	Y	T	temporal hippocampal	IIA	CBZ 1.4k, LV 3k
16	120	38	Y	T	left anterior temporal	IA	LV 2.75k, CBZ 800, PRM 750
17	127	54	Y	A	right frontal temporal insular	IVA	CBZ 1K, CLB 20, LM 200
18	148	25	N	A	left parietal insular	IVA	CBZ 1.2k, CLB 40, FB 75
19	145	62	Y	A	right temporal perysilvian	IIC	FB 150, LS 400, CLB 20
20	159	39	Y	A	right insular perysilvian	IVA	CBZ 800, LM 400

Table 10: Third set of patients. The columns are the same as in Table 8

Subject ID	#C	#PC	P	A/T*	Region*	Engel*	AED* [mg]
21	158	26	Y	-	-	-	-
22	109	13	Y	A	left temporal mesial	IA	CBZ 1.2k, LV 750
23	149	44	N	A	left temporal	IA	LV 3k, CBZ 1k, ls 500
24	79	61	Y	T		IA	CBZ 1.2k, FB 100
25	152	26	Y	A	right frontal	IIA	VP 800, CLB 10
26	149	15	Y	A	right mesial frontal	IIIA	CBZ 800, LV 3k, NTZ 1.5
27	91	11	Y	A	right central frontal	IA	LM 400, LV 2k
28	133	15	Y	A	right frontal	IVA	CBZ 600, RF 1.5k
29	142	15	N	A	right frontal	IA	CBZ 1.2k, ZN 400, FB 1k
30	157	14	Y	-	-	-	-

Table 11: Fourth set of patients. The columns are the same as in Table 8

Subject ID	#C	#PC	P	A/T*	Region*	Engel*	AED* [mg]
31	130	64	N	None	None	None	LV 1.5k, CLB 5
32	139	68	N	-	-	-	-
33	165	51	Y	A	right temporal anterior mesial	IIA	OXC 2k, FB 150
34	137	9	Y	-	-	-	CBZ 1.6k, LV 4k
35	134	68	Y	None	None	None	LV 3k
36	114	61	Y	A	right orbital temporal	IA	ZN 400, LV 750, CBZ 1.4k
37	145	25	N	None	None	None	LS 500, VP 1k, ZN 200
38	101	61	Y	T		IA	CBZ 1k, LV 2.5k
39	127	28	Y	A	left occipital	IIA	CBZ 1.2k, LV 1.5k, LS 300
40	118	38	N	-	-	-	-

Table 12: Fifth set of patients. The columns are the same as in Table 8

Subject ID	#C	#PC	P	A/T*	Region*	Engel*	AED* [mg]
41	130	39	Y	None	None	None	CBZ 1.2k, LV 3k, LS 150, CLB 20
42	140	6	Y	A	left cingulus	IA	OXC 1.8k, TP 200, LV 3k, CLB 10
43	142	22	Y	A	right frontal anterior	IIIA	CBZ 1k, LV 1k
44	139	68	Y	T	right temporal parietal perysilvian	IA	TP 200, LM 200
45	140	53	Y	A	left temporal parietal	IA	CBZ 900
46	143	57	Y	T	right temporal opercular	IVA	CBZ 900, LV 3k
47	149	54	N	-	-	-	-
48	150	76	N	A	left frontal	IA	CBZ 1.2k, LM 200, CLB 20
49	148	34	Y	T		IA	LV 3k, LS 400
50	125	51	Y	A	right temporal occipital	IA	LM 600, LV 2k

Table 13: Sixth set of patients. The columns are the same as in Table 8

Subject ID	#C	#PC	P	A/T*	Region*	Engel*	AED* [mg]
51	144	25	N	A	right temporal	IA	CBZ 1.4k, LV 3k, CLB 10
52	136	36	Y	A	left insular opercular	IIIA	CBZ 800
53	148	25	Y	A	right frontal temporal	IA	CLB 20, LM 600, FN 500
54	88	39	N	A	left temporal insular, opercular	IA	CBZ 1.8k, CLB 20
55	114	15	Y	None	None	None	FN 400, TP 500
56	146	17	Y	A	left temporal	IA	OXC 1.5k, CLB 20, LV 2.5k
57	162	15	N	A	right mesial temporal	IA	TP 75, CBZ 1.5k
58	127	43	Y	T	nodular heterotopia	IVAa	CBZ 1k, LV 500, CLB 20
59	146	12	Y	A	left temporal anterior mesial	IIA	CLB 20, FB 45
60	141	-	-	-	-	-	-

5

Feasibility Study: a Preliminary Approach

In this chapter we report the first attempt at combining machine learning and signal processing for the automatic classification of the epileptic areas. Given the neural recordings, our approach relies on specific preprocessing and feature engineering. These steps are guided by clinical a priori. Once we obtain a representation, we resort to standard learning methods for the classification of the samples. Throughout this analysis we aimed at understanding if the classification of epileptic areas may be approached using standard signal representations of clinical interpretation. This represents a starting point in proposing machine learning and automatization of the analysis as a support to the visual inspection of the data. This analysis has been presented @ CIBB 2018, Lisbon, (Portugal).

Our goal is to exploit standard measures from clinical literature related to the analysis of electrophysiological neural recordings [15]. Given these measures, we address the identification of the EZ combining signal processing and classification techniques. We remark that in this preliminary stage, the patients post-surgical outcome was not available. As such, the evaluation of the pipeline is based on presurgical assessment.

The rest of the chapter is organized as follows: we first give a characterization of the main features extracted from the entire time series, in both the frequency and the temporal domain, and from the area of acquisition. We then describe the main learning methods used for the classification task and we report their performance. Lastly we comment the results and we discuss the limitations of the analysis. The code related to this pipeline is available at `https://github.com/vanessadamario/multichannelAnalysis`.

5.1 Dataset Description

The analysis is performed on the dataset described in Chapter 4. We consider the entire set of 60 patients for which we have different information, as re-

ported in Table 14. The dataset divides in two. For 40 patients, we have access

	contact position	no contact position
# patients	40	20
	Preprocessing	
	Feature Extraction and Learning	Evaluation of Threshold Values

Table 14: The split of the population, given the availability of information related to the contact position. We use the portion which misses the positional information to extract an estimate of the physiological baseline activity.

to the contacts position and the labels related to the presurgical assessment. This set will be used as our learning and test set throughout the analysis. The feature extraction technique will rely nonetheless on extraction of parameters based on standard and physiological activity. We rely on the 20 patients for which we do not have information about the contacts position to extract this information, including the non epileptic subject.

5.2 Feature Engineering

We perform on the entire dataset of 60 patients the same pre-processing procedure. For each time series we remove line effect using a notch filter peaked at 50 Hz and all the harmonics up to the Nyquist frequency, $f_{\text{Nyq}} = 500$ Hz. We use a 2^{nd} order Butterworth filter, bandstop width 2 Hz.

The feature engineering step relies on information extracted from the time series in both the temporal and the frequency domains.

SPLIT OF THE FREQUENCY DOMAIN For most of the extracted features, we divide the spectrum in several frequency bands, shown in Table 15 and Table 16.

Rhythm	*slow*	δ	θ	α	β	γ	high-γ
Frequency [Hz]	< 1	$[1, 4]$	$[4, 8]$	$[8, 13]$	$[13, 30]$	$[30, 70]$	$[70, 90]$
Band name	B0	B1	B2	B3	B4	B5	B6

Table 15: In the first row we report the neurophysiological rhythms, in the second the frequency interval in Hz, and at the last row the assigned name to ease the notation

THRESHOLD VALUES EVALUATION To obtain an average estimate of the standard physiological activity across patients at different frequency bands we considered only bipolar recordings acquired from non-EZ areas, based on pre-surgical assessment.

Frequency [Hz]	[90, 140]	[140, 190]	[190, 240]	[240, 290]
Band name	B7	B8	B9	B10
Frequency [Hz]	[290, 340]	[340, 390]	[390, 440]	[440, 490]
Band name	B11	B12	B13	B14

Table 16: Intervals in which we divide the spectrum at higher frequencies and the high frequency bands names.

Standard activity is evaluated at different frequency bands, as specified in Tables 15 and 16. To filter at a generic band $B*$ we use a band-pass Butterworth filter (2^{nd} order), with cut-off frequencies and central frequency corresponding respectively to the extremes and the center of the frequency interval.

For each non-EZ contact, after filtering at a specific band, we quantify the standard deviation, as $\sigma_{\text{B}k} = \sigma\left[\text{IIR}_{\text{Butter(B}k)}(S(t))\right]$ and its average value across the non-EZ population $\langle\sigma_{\text{B}k}\rangle$. These values are reported in Table 17.

Band name	B0	B1	B2	B3	B4	B5	B6	B7
$\langle\sigma_{\text{B}k}\rangle$ [μV]	17.24	13.19	15.54	12.45	11.09	4.58	1.19	1.07
Band name	B8	B9	B10	B11	B12	B13	B14	
$\langle\sigma_{\text{B}k}\rangle$ [μV]	0.578	0.391	0.298	0.244	0.203	0.183	0.190	

Table 17: $\langle\sigma_{\text{B}k}\rangle$ values for each frequency band extracted from physiological bipolar recordings on the subset of 20 patients. We considered these fluctuations as an estimate of physiological standard activity.

5.2.1 *Features Extraction for solving the Learning Task*

We split the bipolar electrophysiological recordings from the 40 patients (Table 14) in two segments of equivalent length. We operate this choice guided by the strong hypothesis that neural signals during the interictal period, in absence of any external stimulus, may be considered as stationary. This is nonetheless a controversial assumption for electrophysiological recordings, even if operated during the resting state activity and it may affect the analysis [68]. We will observe in the next chapters how correlation of the activity may represent a potential limitation, in terms of overestimation of the classification performance. We furthermore considered one feature inferred from the spatial position of each contact.

Let us enumerate the features extracted for each bipolar segment:

(i) *first moments of the time series*, where the bipolar recordings are considered in the temporal domain and are not filtered at a specific frequency band, but the harmonics of 50 Hz;

(ii) *relative energy*, at the frequency bands specified in Tables 15 and 16;

(iii) *normalized energy from wavelet coefficients*, where instead of the bands the division at different frequency is operated automatically by the mother wavelet at different dyadic scales;

(iv) *wavelet entropy*, which takes into account the decomposition of the signal through the wavelet representation;

(v) *over-threshold activity*, measured as the time spent over an estimation of the average electrophysiological activity, reported for each frequency band in Table 17, we compute also the mean activity and its standard deviation after filtering the signal at each frequency band;

(vi) *partial tissue density*, which is determined by the neural tissue in the area of acquisition.

FIRST MOMENTS OF THE TIME SERIES We consider *variance*, *skewness* and *kurtosis*. We excluded the mean as the time series will be always filtered in band to avoid constant trends.

RELATIVE ENERGY AT FREQUENCY BANDS We measure the energy concentration at the different frequency bands. As in [94], we compute the relative spectrum by normalizing the contribution of the spectrum within the frequency window of interest with respect to the energy for the entire spectrum. Let S be a generic bipolar recording in the temporal domain, we denote its Discrete Fourier Transform as $\hat{S}$. The normalized Energy (nE) at a specific band Bk corresponds to

$$\mathrm{nE}_{\mathrm{B}k}(S) = \frac{\sum_{f \in \mathrm{B}k} \hat{S}^2[f]}{\sum_{g \in [0, f_{\mathrm{Nyq}}]} \hat{S}^2[g]} \tag{51}$$

NORMALIZED ENERGY FROM WAVELET COEFFICIENTS We use the discrete wavelet transform to get the signal decomposition onto an orthogonal basis, with 2nd order Daubechies mother wavelet. Again, we compute the relative wavelet energy at each scale with respect to the total. We measure the energy as the sum of the square of the wavelet detail coefficients, using the wave decomposition function[1]

$$[\mathrm{cA}, \mathrm{cD}] = \mathrm{DWT}(S(t))$$

cA is an array of approximation coefficients, while cD is a list containing the detail coefficients at each scale, using the concept of quadrature mirror filter. The energy at the i-th scale, denoted as WE_i, corresponds to $\mathrm{WE}_i = \sum_j \mathrm{cD}_j^2[i]$, where j is the index denoting the time instant. We quantify the normalized Wavelet Energy at scale i as

$$\mathrm{nWE}_i = \frac{\mathrm{WE}_i}{\mathrm{WE}_{\mathrm{tot}}} \tag{52}$$

as the concentration of the signal at a specific scale.

1 Last access: October 14th, 2019, `https://pywavelets.readthedocs.io/en/latest/ref/dwt-discrete-wavelet-transform.html`

WAVELET ENTROPY This measure has been defined for the analysis of short duration patterns in the signal. It leverages on wavelet transform and it has been shown to be a discriminative quantity in the evaluation of signal coherence in neurophysiology [106], especially in pathological activity detection [88]. Here we will use it naively, by considering the entire time series. The definition of the Wavelet Entropy (WH) is based on the notion of nWE given before. This term quantifies the concentration of the energy at some scales

$$\mathrm{WH}(S) = -\sum_i \mathrm{nWE}_i \log(\mathrm{nWE}_i). \tag{53}$$

OVER-THRESHOLD ACTIVITY As in Bartolomei et al. [11] we quantify the *hyperactivity* of each bipolar recording as an estimate of abnormal amplitudes, with respect to a baseline activity. The threshold may be computed using a segment identified by clinical experts through visual inspection, or may be inferred from some statistics about the distribution of the signal values.

An example of this approach for what regards hyper-activity at high frequency is provided by Staba and collaborators, who make use of 5 times the standard deviation of the root mean square of the signal to identify candidate HFOs [118, 120]. In our case we resort to the values reported in Table 17 as threshold values, which we multiply for positive scalar factors.

Given a bipolar recording we evaluate the presence of over-threshold activity for the frequency bands in Table 17 as follows. Let Bk be the generic frequency band. We band-pass the signal S using a Butterworth 2^{nd} order, with cut-off frequency and central frequency as respectively the extremes and the center of the Bk interval. We fix a constant value C, element of the set $\{2,3,4,5,6,7\}$. We evaluate the time spent by the signal above a threshold, at band Bk, using the results in Table 17

$$\mathrm{Time}_{\mathrm{B}k,C}(S(t)) = \frac{1}{f_s}\sum \mathbb{I}\left[\mathrm{IIR}_{\text{Butter, B}(k)}(S(t)) > C \cdot \langle \sigma_{\mathrm{B}k} \rangle\right], \tag{54}$$

with $\mathbb{I}$ the indicator function and f_s sampling frequency. Moreover, after filtering the signal, we save the mean activity at the standard deviation at the Bk frequency band, for a total of two features per band.

PARTIAL TISSUE DENSITY Despite the great variability across neural populations, the brain tissue can be coarsely divided in *gray matter* and *white matter* (see Chapter 3, [69]). The former has a prevalence of neural cells, glial cells, and vessels. The latter consists mostly of myelinated axons and fibers which connect different brain regions.

The recent work of Mercier [83] gives a broad insight in the role played by white matter in the signal propagation. Mercier highlights the improvement obtained by adopting the bipolar montage, which allows to decouple spurious activity. His work also shows the relevance of quantifying the anatomical nature of the brain tissue in the area of acquisition of the signal. Indeed in the same work, the electrode position is proved to be crucial for clinical evaluations, as there is a high correlation between signal power and the presence of gray matter over white matter regions.

In this analysis the differentiation between gray and white matter is possible using Freesurfer [42], a software tool which parcellates cortical and subcortical regions from MRI acquisition. The Partial Tissue Density (PTD) index is defined as

$$\text{PTD} = \frac{\text{Vox Gray} - \text{Vox White}}{\text{Vox Gray} + \text{Vox White}}, \tag{55}$$

where Vox Gray and Vox White correspond respectively to the number of gray and white voxels contained in a volume of 3mm^3 centered around the electrode position [83]. Note that with PTD $\in [-1, 1]$. We extract the PTD index for each contact in the monopolar setting. We assign to each bipolar recording the PTD index of the first contact between the two, the deeper in the brain tissue.

5.2.2 *Classification*

We concatenated the features extracted at the previous step in a unique vector. The generic sample of our dataset (X_i, y_i), with $X_i \in \mathbb{R}^{156}$ collecting the extracted features and $y_i \in \{-1, 1\}$ binary label, which assesses the epileptogenicity for the recorded area. The proportion of epileptogenic and non epileptogenic channels is unbalanced in favor of non epileptogenic contacts, with random guess corresponding to 0.74%.

In Table 18 we report the main aspects characterizing our dataset.

Features	moments	energy	wavelet	entropy	threshold	PTD
	3	15	16	1	120	1
Data	non EZ for thr		EZ for clas		non EZ for clas	
	1968		1342		3973	

Table 18: Summary of the dataset used for this experiment. In the first row we report the subsets of features, which mostly are extracted in the temporal domain, but one related to the spatial position of the contact, for a total of 156. The bottom row contains the number of contacts used respectively to compute the threshold values (non EZ for thr), the number of EZ contacts (EZ for clas), and non EZ contacts (non EZ for clas) in the classification task.

We consider several binary classification techniques, both linear and non-linear: Logistic Regression (LR), SVM with linear kernel, Random Forest (RF) and Gradient Boosting (GB).

In LR we imposed sparsity through the L^1 norm, with the regularization constant C varying in a logarithmically spaced range of twenty values between $(10^{-2}, 10^2)$. For SVM, we fixed a linear kernel and let the cross validation choose the best values of C, in the same range of LR. For what concerns RF we fixed the number of estimators to 10^3 where the tunable parameters were the percentage of maximum features with respect to the total, in the range $(0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$. In GB we fixed the learning rate to 10^{-3}, the tunable parameters were the max depth of trees, free to vary linearly in a range be-

tween $(3, 31)$ and the number of estimators, chosen between three linearly spaced values in $(100, 500)$.

We split the dataset in 85% samples for learning and 15% for test, using automatic `scikit-learn` [98] procedures that split the dataset with respect to the unbalance of the original problem. The choice of the optimal hyper-parameters for all the algorithms was performed by using 3-fold cross-validation in the learning procedure.

5.2.3 *Results*

The learning and testing procedure has been repeated 50 times using the Monte-Carlo strategy in order to get a statistically reliable outcome for the four classifiers, which we report in Table 19. We computed the performance of our methods using metric scores that take into account the unbalance of the dataset, such as precision, recall, balanced accuracy (Balanced Acc), and F1 score.

Classifier	Precision	Recall	Balanced Acc	F1 score
LR	0.68 ± 0.03	0.34 ± 0.03	0.64 ± 0.01	0.45 ± 0.03
SVM	0.57 ± 0.04	0.15 ± 0.03	0.55 ± 0.01	0.23 ± 0.05
RF	0.88 ± 0.02	0.57 ± 0.02	0.77 ± 0.01	0.69 ± 0.01
GB	0.80 ± 0.07	0.35 ± 0.06	0.66 ± 0.02	0.48 ± 0.05

Table 19: Average classification performance obtained across 50 repetitions of the experiment. Random forest obtains the best predictive performance across all the metrics.

RF obtained overall the best performance across all the metrics. All metrics show a performance which is highly above chance level, which is promising in the discrimination of epileptic areas. The precision value for random forest indicates that the number of false positives is relatively low.

5.3 Comments

The integration of spectral features with anatomical characteristics of the recorded areas for posterior localization based on MRI test is a first timid attempt to merge multiple clinical tests and to fix a set of features which are both functional and structural descriptors of the epileptic brain.

Nonetheless, there are several flaws relative to the learning pipeline and the feature extraction part. Firstly, the split of the recordings in temporal windows of five minutes length has been made to augment the dimensionality of the dataset, in the hypothesis that the interictal period should not contain any causal relation. The presence of temporal correlations across samples arising from this decision has not been verified in this analysis and cannot be excluded.

Another critical aspect is related to the mix of all recordings in the learning and testing procedure. We proceeded to the feature extraction step and we put together the features extracted from all the samples, without considering the possibility of having recordings from the same patient in different splits. This decision is not an optimal experimental choice, as in terms of predictive tools the results may suffer from the presence of correlations. We will address all these potential issues in Chapter 7.

6

Interpretable Decision Support Tool through Data Integration: Multi Task Multiple Kernel Learning

In this chapter we present one of the main contributions of this thesis, which integrates non-homogeneous SEEG recordings derived from multiple patients. As we have seen, the number of contacts changes across the population, not allowing a direct comparison among patients. The differences due to SEEG montages will be overcome through a multi-task functional, where each patient represents a task. The outcome of the method incorporates both a personalized description for each patient and the selection of the best descriptors of the pathology across the population. The method has been thought as a supporting tool for the identification of pathological recordings. The clinicians would pre-evaluate a subset of signals in the dataset to get a classification for the ones which still have not been analyzed. The results have been presented @Invasive Mathematics 2018, Genova (Italy) *and* @ MLHC 2018, Stanford, CA (USA). *The method received several key comments and observations which will inspire further investigations, throughout this chapter and the entire work. In this chapter we test the algorithm in different conditions, in order to evaluate its strength and limitations.*

6.1 Goals and Contribution

Our main goal is to support medical experts in the diagnosis of the focal areas, by giving also interpretable results. To this aim we design Multi-Task Multiple Kernel Learning (MT-MKL), a tool for the analysis of SEEG recordings. MT-MKL is a supervised method which fuses the multi-scale representation of time series and feature selection techniques to understand which features play the major role in the predictive task across multiple patients.

Through the *multiple kernel* component of the algorithm we aim at identifying the relevant information in terms of amplitude and phase similarities at certain frequencies. We argue that comparing the signals by first decomposing their contributions in different frequency bands is crucial. To illustrate this we report in Figure 21 an example of a spectrum ($\mathcal{F}(S)(f)$) of a bipolar

recording at the y-log scale. The amplitude of the signal decays as the fre-

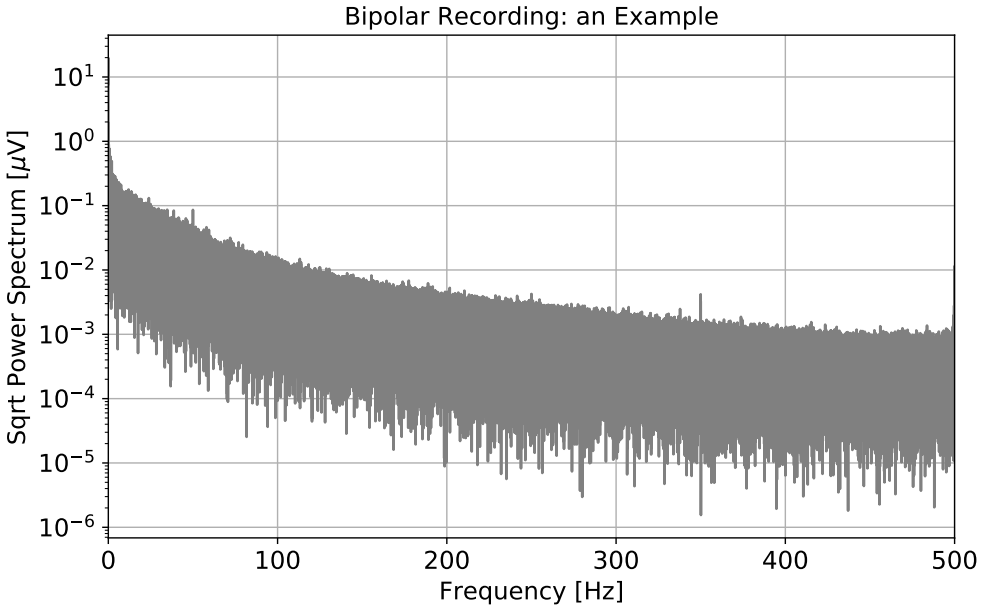

Figure 21: We report here the spectrum of a pathological bipolar recording, from class Engel I. We observe that the power distribution goes to zero as the frequency increases. The neurophysiological signal follow the power law $1/f^{\beta}$.

quency increases. From [53, 86] there is common agreement on the fact that the neurophysiological signal follows a power law $1/f^{\beta}$.

In order to get a fair comparison of different frequency bands in the predictive task the idea here is to split their contributions. The multi-scale decomposition of the signal was inspired by the work of Mallat [80]. Mallat analyzes the properties of the scattered transforms, which resemble Deep Convolutional Neural Networks (DCNN). Differently from the DCNN, we fix the filters as mother wavelets with varying scales.

We leverage on a shallow wavelet representation, and on top of this we apply feature selection. The wavelet scales and the similarity metrics are shared across the entire population.

The *multi-task* nature of the method arises from the need of merging information and finding out relevant epileptogenic features, while classifying epileptogenic regions with such different SEEG montages. Indeed the classification of SEEG recordings for each patient constitutes a single learning task. In this way, our model is the sum of personalized models built for each patient, where the features are nonetheless shared across the entire population. We performed the selection of relevant scales through a feature selection approach, by imposing sparsity on the coefficients related to the similarity metrics. Sparsity constraints are applied also on single models, for the aim of using a small amount of bipolar contacts, so to observe which are the most contributive to the classification task.

MT-MKL is not totally automatic as it cannot be used for predictive task on a new set of patients. Given a patient we will compare, at each scale, the representations of the recordings in the training set, giving a measure of their similarity. The classification of an unseen bipolar recording from p relies indeed on the pairwise comparison with recordings from the training set. This explains the need of having for each patient some labeled samples. Nonetheless, differently from the analysis presented before, MT-MKL takes into account the SEEG montage, including the variety in the number of contacts, which strongly depend on the clinical non invasive assessment about the candidate epileptogenic areas.

In Section 6.2 we introduce the wavelet representation and the similarity metrics we leverage on to solve the classification task. In Section 6.3 we consider the optimization problem. In Section 6.4 we present the experimental setup and the main results obtained from the MT-MKL. The method has been presented @ Invasive Mathematics (Genova, May 2018), @ Machine Learning for HealthCare conference (Stanford, August 2018), and @EuroScipy 2018 (Trento, September 2018). These three events were extremely useful for comments that guided us to further changes and analysis, which will be widely discussed in Section 6.5.

6.2 Data Representation and Similarity Measures

We denote a generic i-th bipolar recording from a patient p as $S_i^{(p)}$. To ease the notation, we omit the patient index from now on. Please note that the similarities are always computed on a generic pair of bipolar recordings $\left(S_i^{(p)}, S_j^{(p)}\right)$ coming from the same p patient, even if we omit the notation.

6.2.1 *Multi-Scale Representation of the Recordings*

The multi-scale representation for S_i is obtained through the CWT, introduced in Eq. 8. We use the complex Morlet transform as mother wavelet

$$\Psi(t) = \frac{1}{\sqrt{\pi}B} \exp\left[2i\pi tC\right] \exp\left[-\frac{t^2}{B}\right], \tag{56}$$

with $B = 1\ \mathrm{s}^2$ and $C = 1\ \mathrm{s}^{-1}$ for the default sampling period of 1 second. The complex part of this transformation is such that it enable us to capture aspects related to both the amplitude of the signal as well as its phase.

The transformation as in Eq 57 will be used across the experiments.

$$\Psi_{\tau,s}(t) = \frac{1}{\sqrt{\pi s}} \exp\left[2i\pi\frac{t-\tau}{s}\right] \exp\left[-\frac{(t-\tau)^2}{s^2}\right], \tag{57}$$

with the tuple (τ, s) denoting respectively the temporal shift and the scaling parameter. We define the wavelet representation of the signal as

$$(\mathcal{W}S_i)(\tau, s) = \frac{1}{\sqrt{s}} \int d\xi S_i(\xi) \Psi\left(\frac{\xi-\tau}{s}\right). \tag{58}$$

Let $A_i(\tau, s)$ be the instantaneous *amplitude* at scale s of the wavelet representation of the signal

$$A_i(\tau, s) = |(\mathcal{W}S_i)(\tau, s)|. \tag{59}$$

Let $\Phi_i(\tau, s)$ be the instantaneous phase at time τ and s scale of the representation

$$\Phi_i(\tau, s) = \arctan\left[\frac{\Im[(\mathcal{W}S_i)(\tau, s)]}{\Re[(\mathcal{W}S_i)(\tau, s)]}\right], \tag{60}$$

with $\Im$ the imaginary part and $\Re$ the real part of the wavelet coefficients.

6.2.2 *Similarity Measures*

To measure the similarity of two time series, we consider standard measures of pairwise correlations in phase, amplitude, and in the frequency domain. In the first case we resort to the Phase Locking Value (PLV), which quantifies the coherence at each time point. In the second case we use a normalized correlation to evaluate the amplitude similarity. This quantity is nonetheless not invariant to temporal shifts. Correlation is a reliable quantity when the lag between similar patterns is negligible compared to the pattern length, e.g. areas participating to the same neural activity or almost instantaneous propagation. Finally to capture similar behavior of two recordings, independently from the temporal lag between the two, we resort to shift invariant spectral measures.

PHASE LOCKING VALUE A measure of phase synchrony between bivariate measures is known as Phase Locking Value (PLV) [72]. We define the PLV at scale s as

$$\mathrm{PLV}_s(S_i, S_j) = \frac{1}{T}\left|\sum_{\tau=1}^{T} \exp\Big[-i\big(\Phi_i(\tau, s) - \Phi_j(\tau, s)\big)\Big]\right|, \tag{61}$$

with T length of the recordings. Its value ranges between $[0, 1]$. The maximum PLV value corresponds to a pair of perfectly synchronous signals.

NORMALIZED CORRELATION Let $\mu_s(A_i) = \mathbb{E}_\tau[A_i(\tau, s)]$ be the empirical expectation value of the wavelet coefficients amplitude at scale s. Let $\sigma_s^2(A_i) = \mathbb{E}_\tau[(A_i(\tau, s) - \mu_s(A_i))^T(A_i(\tau, s) - \mu_s(A_i))]$. We write the covariance at scale s as

$$\mathrm{cov}_s(S_i, S_j) = \frac{\mathbb{E}_\tau\Big[\big(A_i(\tau, s) - \mu_s(A_i)\big)^T\big(A_j(\tau, s) - \mu_s(A_j)\big)\Big]}{\sigma_s(A_i)\sigma_s(A_j)}. \tag{62}$$

SPECTRAL MEASURES We define the cross power spectral density as the Fourier transform of the convolutional product, denoted by $*$, of the absolute value of wavelet coefficients for the two signals, computed as follows

$$P_s(S_i, S_j)(f) = \mathcal{F}(A_i * A_j)(f). \tag{63}$$

The module of cross power spectral density is invariant in the time and frequency domain, as a result of the Parseval theorem [124]. In order to quantify the similarity between spectra, we normalize the quantity in Eq. 63

$$\tilde{P}_s(S_i, S_j) = \frac{\|P_s(S_i, S_j)\|^2}{|P_s(S_i, S_i)| \cdot |P_s(S_j, S_j)|}. \tag{64}$$

6.3 Learning Method and Feature Selection

We refer to the entire dataset as $\mathcal{D} = \{S^{(p)}, y^{(p)}\}_{p=1}^{N}$ with N number of patients. For the generic patient p, the number of recordings c^p varies and consequently the dimensions, so that $S^{(p)} \in \mathbb{R}^{c^p \times N_T}$ and $y^{(p)} \in \{-1, 1\}^{c^p}$. The term N_T denotes the number of time points.

6.3.1 *Multi Task Multiple Kernel Learning*

Multiple Kernel Learning (MKL) [16, 73] integrates data by combining different kernel functions. A straightforward MKL may be a linear combination of different kernels [16]. This is possible given the fact that kernels allow linear operations while preserving their mathematical properties, e.g. positive semi-definite and symmetry [52].

Kernels K are positive semi-definite matrices whose entries K_{ij} encode pairwise similarity between the i-th and j-th samples (Chapter 3, [113]). The metrics choice, which reflects in the choice of a suitable kernel function can be tricky as it heavily depends on the data at hand and the task we aim at solving.

Formally, consider k kernels $\{K_1, \ldots, K_k\}$ that represent different similarity measures among dataset samples. Such kernels can be combined linearly as a weighted sum $\sum_{i=1}^{k} w_i K_i$, where $w = (w_1, \ldots, w_k)^\top \in \mathbb{R}_+^k$ is a vector of non-negative components, obtained through an optimization procedure. Such components measure the importance of each kernel for the problem at hand.

MKL in its general formulation is not suitable to our dataset [45]. Indeed, in our case the variability of the implantation setting does not allow for a direct comparison of the neural activity across patients. In other words, it is not possible to use a unified regression model for all patients, as we need to look at a portion of the contacts for each patient in order to classify the ones left out.

To handle the variability problem we extended the MKL to Multi-Task Multiple Kernel Learning to account for different patient conditions. Each kernel represents a particular similarity matrix among all the contacts in a single patient at a specific scale. The innovation in the method consists in the capability of jointly analyzing the patients activity by taking into account their diversity.

We denote $\left\{K_1^{(p)}(\cdot), \ldots, K_k^{(p)}(\cdot)\right\}$ the set of k similarity maps for the generic patient p. The decision function $f^{(p)}$ for the patient p and the recording S_q is defined as

$$f^{(p)}\left(S_q^{(p)}\right) = \alpha_0^{(p)} + \sum_{i \in \mathcal{C}_p^{\text{tr}}} \left[\alpha_i^{(p)} \sum_{j=1}^{k} w_j K_j^{(p)}\left(S_q^{(p)}, S_i^{(p)}\right)\right], \tag{65}$$

with $\alpha_i^{(p)}$ the i-th component of the regression parameter $\alpha^{(p)}$, specific for each patient p, $\mathcal{C}_p^{\text{tr}}$ set of contacts used during training.

Having separate parameters $\left(\left\{\alpha^{(1)}, \ldots, \alpha^{(N)}\right\}, w\right)$ is fundamental for the resolution of our problem. In fact, $\alpha^{(p)}$ allows to better approximate the labels $y^{(p)}$ by capturing the variance of each patient, while w combines the kernels by weighting them and, as it holds across patients, it provides relevant indication of the most discriminative kernels.

In order to obtain interpretable results and a more stable solution we also add an elastic-net (or L^1L^2) penalty on w and $\alpha = \left\{\alpha^{(1)}, \ldots, \alpha^{(N)}\right\}$. By considering all the patients, our goal translates into minimizing the following objective function:

$$\begin{aligned} \underset{\alpha^{(1)},\ldots,\alpha^{(N)},w}{\text{minimize}} \Bigg\{ & \sum_{p=1}^{N} \left(\ell_{f^{(p)}}\left(S^{(p)}, y^{(p)}\right) + \lambda(r_\lambda \|\alpha^{(p)}\|_1 + (1 - r_\lambda)\|\alpha^{(p)}\|_2^2)\right) \\ & + N\beta\left(r_\beta \|w\|_1 + (1 - r_\beta)\|w\|_2^2\right) \Bigg\} \\ & \text{s.t. } w_j \geq 0 \text{ for each } j = 1, \ldots, k. \end{aligned} \tag{66}$$

The single-task loss function is the negative log-likelihood of the logistic probability function

$$\ell_{f^{(p)}}\left(S^{(p)}, y^{(p)}\right) = -\sum_{i \in \mathcal{C}_p^{\text{tr}}} \log\left[1 + \exp\left[-y_i^{(p)} f^{(p)}\left(S_i^{(p)}\right)\right]\right]. \tag{67}$$

The terms r_λ and r_β are hyper-parameters of the elastic-net penalty ratios on α and w respectively. The elastic-net penalty benefits indeed from the well-known stability property of the L^2 regularization term [138].

6.3.2 *Minimization*

For the optimization of the functional in Eq. 66 we rely on alternating minimization [14]. Indeed after a revision on the current state-of-the-art of multiple kernel methods [45] we did not assume our problem to be jointly convex in both w and the set of $\left\{\alpha^{(1)}, \ldots, \alpha^{(N)}\right\}$ weights. Despite this is one of the main features characterizing classical MKL methods, the adopted optimization framework [14] guarantees convergence to a critical point under mild assumptions of non global Lipschitz continuity for the gradient of the smooth term.

In particular the optimization of the form in Eq. 66 is based on an alternating forward-backward splitting procedure given the non-differentiability of some parts of the functional (L^1 norm) [14, 27]. The optimization procedure is described in Algorithm 1.

Algorithm 1 Alternating Minimization Algorithm

1: Initialize $\left(\left\{\alpha^{(1)}(0), \ldots, \alpha^{(N)}(0)\right\}, w(0)\right)$
2: **for** $t < t_{\max}$ **do**
3: **for** $p = 1, \ldots, N$ **do**
4: $\alpha^{(p)}(t) \leftarrow$ minimize Problem 66 with $w = w(t-1)$
5: **end for**
6: $w(t) \leftarrow$ minimize Problem 66 with $\alpha = \alpha(t)$
7: **if** stop criterion is met **then return** $\left(\left\{\alpha^{(1)}(t), \ldots, \alpha^{(N)}(t)\right\}, w(t)\right)$
8: **end if**
9: **end for**

MINIMIZATION OF α. Fixing w, for each patient p the functional w.r.t. $\alpha^{(p)}$ takes the form of a standard logistic regression. Its minimization is then performed by computing the derivative on the logistic loss and the L^2 norm and then applying the soft-thresholding operator [96] on the result of the gradient descent step. The gradient for the q-th component of the vector $\alpha^{(p)}$ relative to the patient p is

$$\nabla \alpha_q^{(p)} = \frac{y_q^{(p)} \exp\left[-y_q^{(p)} f^{(p)}\left(S_q^{(p)}\right)\right]}{1 + \exp\left[-y_q^{(p)} f^{(p)}\left(S_q^{(p)}\right)\right]} \sum_{j=1}^{k} w_j K_j^{(p)}\left(S_q^{(p)}, \cdot\right) + 2\lambda(1 - r_\lambda)\alpha_q^{(p)}. \quad (68)$$

MINIMIZATION OF w. The minimization of w is more tricky, given its non-separability across various patients. In Eq. 69, we report the gradient of the differentiable part of the functional, which consists in the sum of gradient computed for each patient p and the L^2 term. Then, we apply the soft-thresholding operator to enforce sparsity in the solution. Also, we project the kernel weights into the positive half-space by applying a threshold on zero. This ensures that a kernel is considered only if its weight is positive, otherwise it is discarded. The gradient for the generic m-th component of the kernel weight vector is

$$\nabla w_m = \sum_{p=1}^{N} \left(\sum_{i \in \mathcal{C}_{\text{tr}}^{(p)}} \frac{y_i^{(p)} \exp\left[-y_i^{(p)} f^{(p)}\left(S_i^{(p)}\right)\right]}{1 + \exp\left[-y_i^{(p)} f^{(p)}\left(S_i^{(p)}\right)\right]} \left[\alpha_i^{(p)} K_m^{(p)}\left(S_i^{(p)}, \cdot\right)\right] \right) + 2N\beta(1 - r_\beta) w_m. \quad (69)$$

6.3.3 *Parameters Choice*

The choice of the optimal parameter for the model leverages on a k-fold Cross Validation (CV). The best $\left(\beta^*,\ \lambda^*,\ r_\lambda^*,\ r_\beta^*\right)$ are chosen based on the highest averaged balanced accuracy score.

6.4 Experiment I: Support Tool and Feature Extraction

Since MT-MKL needs in input tagged recordings from the positive and negative classes for each patient, we decide to consider only patients who present at least 25% recordings tagged as epileptogenic. We analyze in total 18 patients, for a number of 2347 bipolar recordings, of which 984 have positive label, reported in Table 20.

ID	4	15	17	20	24	31	33	36	38
Engel class	I	II	IV	IV	I	-	II	I	I
#PC /#C	0.41	0.37	0.46	0.25	0.77	0.49	0.31	0.53	0.60
ID	41	44	45	46	47	48	50	54	58
Engel class	-	I	I	IV	-	I	I	I	IV
#PC /#C	0.30	0.49	0.38	0.40	0.36	0.51	0.41	0.44	0.34

Table 20: Experiment I: dataset used at the first implementation of MT-MKL. We stratified the patients, requiring the amount of positive contacts over the total #PC/#C to be greater than 0.25. At the time of the experiment, the post-surgical outcome was not available.

We report in Figure 22 a schematic representation of the implementation of Experiment I.

We first execute the signal processing step, on the top-left, by removing power-line effects (50 Hz and harmonics) using a stop-band 2$^{\text{nd}}$ order Butterworth filter, with 2 Hz bandwidth and filtfilt option to avoid phase distortion. We reduce each recording to a shared length of $T = 590$ s, or $N_T = 590000$ time points. We transform each bipolar recording as in Eq. 58. The shift parameter τ takes discrete values in $[0, N_T - 1]$. The array of wavelet scales is fixed to be a list of one hundred elements equally spaced in the logarithmic scale in the interval $[0.3, 3]$. Fixing s, the central frequency f_a of the mother wavelet corresponds to $f_a = 1/s \cdot t_s$; with t_s denoting the sampling period, equivalent to 1 millisecond. Consequently, the values of f_a vary in the range between 0.5 Hz and the Nyquist frequency, corresponding to 500 Hz. We rely for this step on the MATLAB implementation of the complex Morlet transform [1].

After the preprocessing phase, middle box in Figure 22, we provide the multi-scale representation as input to the algorithm which computes normal-

1 https://www.mathworks.com/help/wavelet/ref/cwtold.html

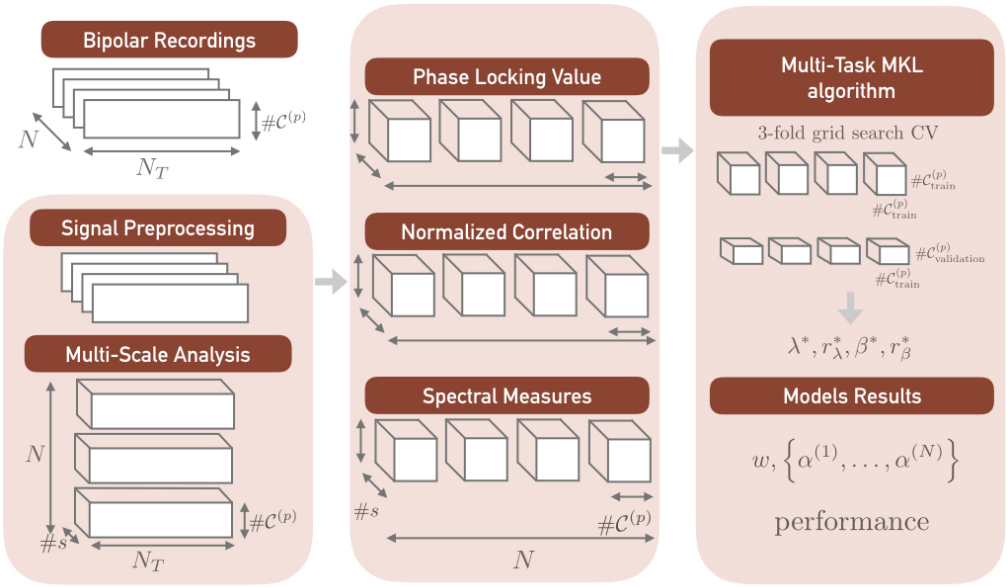

Figure 22: Schematic representation of the learning pipeline of Experiment I. From top left, SEEG recordings are preprocessed to eliminate line contributions. For the multi-scale analysis, we use CWT to represent the time series. The central panel represents the similarity measure computation step, applied for each scale of the wavelet transform. We have in total $s \times 3$ similarity measures. In the last panel, the MT-MKL algorithm includes the minimization and the choice of the best model. MT-MKL returns the set of kernels' weights, the contacts weights, and the predictive result from the logistic probability function.

ized correlation, PLV, and spectral measures. Then, for each patient p, during the training procedure we extract randomly, by respecting the proportion between the two classes, a set $\mathcal{C}_{tr}^{(p)}$ data which are transformed into $k = 3 \times 100$ kernels, each of dimension $\#\mathcal{C}_{tr}^{(p)} \times \#\mathcal{C}_{tr}^{(p)}$. Since quantifying spectral similarity in Eq. 64 is computationally expensive, given the high number of time points for each time series, we approximated this quantity by averaging its estimation on smaller, non-overlapping windows of the signal (5.9 seconds length each).

Lastly, right box in Figure 22, we apply MT-MKL on the resulting similarity measures. The split in training and validation sets is performed by dividing equally the number of recordings (50%/50%) for each patient, while preserving the ratio between the two classes. The learning set is then used to select the optimal hyper-parameters with a 3-fold grid search CV and the score is computed on the validation set. We repeat this procedure 50 times in order to assess the performance stability.

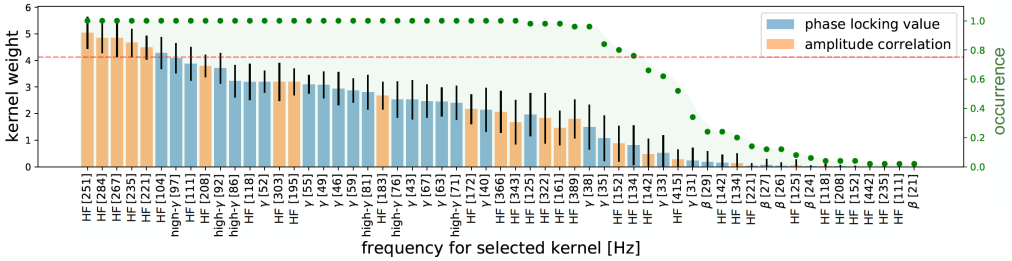

Figure 23: Kernels which mostly contribute in the characterization of the epileptogenic areas. These measures are reported on the x-axis. In square bracket we put the central frequency values of the mother wavelet, and the typical event type related to each frequency. We assign blue color to phase measures and orange to amplitude. Each bar and black line correspond respectively to the mean value and standard deviation of the weights across 50 repetitions of the experiment. The right y-axis denotes the occurrence value, the green dots correspond to the number of times each kernel was selected throughout the repetitions. The dashed line indicates the 0.75% of occurrence value.

6.4.1 *Results*

First we analyzed w, whose components weight each similarity measure, shared across all patients. The L^1L^2 penalty in Eq. 66 gives a sparse small-normed vector w. The non-zero components of this vector can be analyzed to extract information about the importance of similarity measures at specific frequency bands to the prediction task. In Figure 23 we show the similarity measures selected at least once and ordered by their occurrence across 50 repetitions. The most representative similarity measures ($\geq$ 75% occurrence) and the related central frequencies are reported in Table 21 ordered by the mean value of their coefficient.

Normalized Correlation									
Event Type	HF	HF	HF	HF	HF	HF	HF	HF	HF
Central Frequency[Hz]	251	267	221	303	183	366	322	389	134
Phase Locking Value									
Event Type	HF	HF	hγ	hγ	hγ	hγ	hγ	HF	γ
Central Frequency [Hz]	104	111	86	52	76	67	71	125	35

Table 21: On top: most relevant frequencies for the characterization of critic areas related to the signal amplitude. High Frequencies (HF) emerge as the most predictive. On bottom: most relevant frequencies emerging from phase similarity. There is a strong prevalence of the γ and high-γ (hγ) frequencies. Phase Locking Value influences the prediction at lower frequencies than Normalized Correlation.

We notice that the greatest components of w correspond to amplitude correlation at high frequency and phase synchrony at γ and high-γ bands. Note

that the learning pipeline never selects the spectral measures as reliable features for prediction, across all repetitions of the experiment. At the time of the experiments we retained the selection of relevant similarity measures at specific frequency bands to constitute the most statistical reliable payback of the entire procedure, as it is computed across patients.

Another MT-MKL outcome includes statistics on the set of coefficients $\{\alpha^{(1)}, \ldots, \alpha^{(N)}\}$ specific of each patients. These coefficients weight the contacts which have been mostly selected across repetitions and could be considered as the most useful to the classification task. We report in Figure 24 the α vector of a patient from class Engel I.

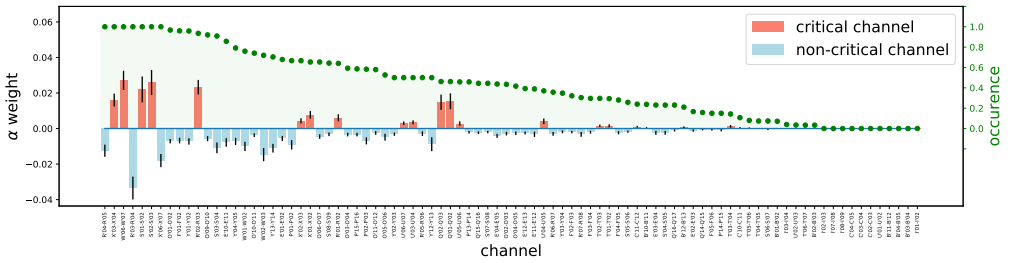

Figure 24: α weights evaluated on a patient. The bars denote the mean weight for each contact, the black bars are standard deviation values across the 50 repetitions. Cyan and red colors refer to the two classes. The green y-axis report the normalized time in which the contact has been selected with the L^1L^2 norm.

We show in Figure 25 the mean and standard deviation of metrics obtained across 50 repetitions of the experiment on the 18 patients. We measured the performance of our model according to the following metrics scores: Precision (P), Recall (R), True Negative Rate (TNR), False Positive Rate (FPR), False Negative Rate (FNR), F1 score, and Balanced Accuracy (BA).

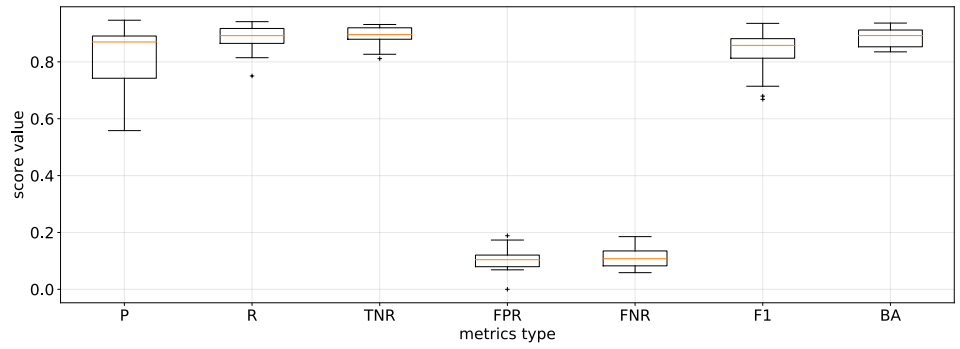

Figure 25: Average of performance scores across patients. Mean and standard deviation are computed across 50 repetitions for each patient over all bipolar recordings from the validation set.

6.5 Comments on Experiment I

This first implementation of MT-MKL has been presented first as a poster at Invasive Mathematics[2] (Genova, May 2018), where I mostly interacted with Christian Benar and Lino Nobili. The work had been previously submitted at MLHC 2018 (April 2018) and accepted in its extended version as a conference paper. The work received several comments from the attendees of the MLHC conference[3] (Stanford (CA), August 2018) and we obtained further feed-back about the pipeline implementation at the EuroScipy 2018[4] (Trento, September 2018). These interactions have been extremely useful as they put in light some potential issues and further experimental setups which should have been tested to improve the analysis.

In the light of this feedback, the following subsections are key as they will guide the analysis contained in this chapter and throughout the overall work.

6.5.1 *Unified Implementation and Filter Design*

In Experiment I, the multi-scale representation and the computation of similarity measures represented the main bottleneck of the entire procedure. In Experiment I MT-MKL relies indeed on both the use of MATLAB and Python software. We leveraged on the former for signal preprocessing, wavelet transforms, and the computation of the cross spectral density measures. The learning pipeline stands instead on top of the scikit-learn [99] implementation of logistic regression, to solve a multitask problem and it was developed in Python. Nonetheless an entirely open-source library would have been much more desirable. In this regard a new release of PyWavelets, with wide implementation of the Continuous Wavelet Transform, and several software tests for a straightforward comparison with the MATLAB software[5] was available. A more formal characterization of the filter width was also missing from our work at this point. The central frequency individuates indeed the main frequency involved but does not clearly quantify the filter width in the frequency domain.

We address these points in Section 6.6, Experiment II. We will provide a new implementation of the learning pipeline entirely based on Python software, with detailed characterization of the filters choice.

6.5.2 *Discrimination of Pathological Rhythms and Patterns*

Frequency bands may not represent the best criterion to identify the pathological contributions of specific patterns to the signal. High frequency bands may collect contributions not only from HFOs, but other patterns, e.g. ISs. The presence of high frequencies (above γ) as a hypothesis for the HFOs contribu-

2 Last visit: November 5th, 2019 `http://mida.dima.unige.it/invasive_mathematics/`
3 Last visit: October 17th, 2019 `https://www.mlforhc.org/2018`
4 Last visit: November 5th, 2019 `https://www.euroscipy.org/2018/program.html`
5 `https://pywavelets.readthedocs.io/en/latest/`

tion cannot be corroborated without a previous localization of interictal spikes and the evaluation of their contribution to the filtered signal. This observation, besides being extremely useful, represents a critical aspect. The automatic interpretation of waveforms in the neural signal would require the use of sophisticated strategies for classifying the patterns. In this regard, the work of Roehri et al. [105] has been suggested as a starting point for an easier separation of HFOs and ISs at high frequencies.

Quantifying the importance of short bursts of activity during the interictal stage will be at the center of Chapter 7 and will lead us to the analysis in Chapter 8.

6.5.3 *Spurious Spatial Correlations*

A further concern is related to the plausible presence of spurious correlation, which could lead to biased classification performance.

This hypothesis is corroborated by the absence of spectral similarity measures as relevant features for classification, across all frequencies. Indeed, we would have expected similar contributions from cross correlations as well as from amplitude similarity. This enforces the possibility that physiological activity together with the effect of spatial correlation may play a major role in the predictive performance.

To analyze if spurious effects plays a substantial role in prediction, we carry out Experiment III, in Section 6.7. We split the recordings in the temporal domain to assess the stationarity in the epileptic behavior. Nonetheless in this setting we are more interested in studying if the prediction remains highly above chance, and if this happens also when we permute the labels. The conclusions obtained from this Chapter will guide us to the experimental design of Chapter 7. There we will test if any possible contribution to the predictive model derives from spurious physiological activity.

6.5.4 *Surgical Outcome and Clinical Validation*

All the considerations were based at the time on the pre-surgical assessment only. Missing post-surgical outcomes impairs the significance of these results. In Table 20 we report the post-surgical outcome which was not available at the time of the analysis. The analysis based on post-surgical outcome will follow in Chapter 7.

6.5.5 *Algorithmic Issues*

At the first implementation the functional in Eq. 66 was normalized by the number of N patients, but in Eq. 65 the normalization factor given by the amount of training contacts per patient $\#\mathcal{C}_p^{\text{tr}}$ is missing. We introduce this factor from Experiment II.

The model relies moreover on the choice a discrete amount of hyper-parameters and several intrinsic parameters. As defined in Eq. 66 MT-MKL is not convex in the tuple $(\{\alpha^{(1)}, \ldots, \alpha^{(N)}\}, w)$. This leads to optimization issues, as convergence to a global minimum is not guaranteed. Several empirical strategies could be implemented to attenuate these effect, as re-training the models using multiple initializations and the use of accelerated minimization methods, which may reduce the chance of converging at local minima with poor generalization.

6.6 Experiment II: Improvements and Usage over all Population

Here we address the issues described in Section 6.5.1. We furthermore fix the normalization of the discriminative function discussed in 6.5.5.

We design a cascade of consecutive wavelet filters, where the number of scales is imposed by requirements on the filters width and their superposition. Overlapped filters should limit information loss for the signal representation. At the same time, L^1L^2 penalty in the functional should protect from instability which rises from sparsity requirements in presence of correlated features.

6.6.1 *Analysis of Morlet Wavelet*

We analyze the effect of the (B, C) parameters on the filters shapes, starting from Eq. 56. Indeed, these are independent from the scale and shift parameters (τ, s). Given that the complex Morlet wavelet is a sinusoidal function modulated by a Gaussian envelope, the (B, C) terms regulate respectively the variance of the Gaussian and the sinusoid frequency. The Fourier transform for the function in Eq. 56 corresponds to

$$\begin{aligned}
\mathcal{F}(\Psi)(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\pi B}} \exp\left[-\frac{t^2}{B}\right] \exp\left[i2\pi Ct\right] \exp[-i\omega t] dt \\
&= \frac{1}{\sqrt{2\pi^2 B}} \int_{-\infty}^{+\infty} \exp\left[-\frac{t^2}{B}\right] \exp\left[-2it\frac{\omega - 2\pi C}{2}\right] dt \\
&= \frac{1}{\sqrt{2\pi^2 B}} \exp\left[-\frac{(\omega - 2\pi C)^2 B^2}{4}\right] \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{B}\left(t + i\frac{\omega - 2\pi C}{2} B\right)^2\right] dt.
\end{aligned}$$

The second last integral can be solved analytically [6] with the following result

$$\mathcal{F}(\Psi)(\omega) = \frac{1}{\sqrt{2\pi^2 B}} \exp\left[-\frac{(\omega - 2\pi C)^2 B^2}{4}\right] \sqrt{\pi B}. \tag{70}$$

With the factorization at the exponent, the function in the frequency domain corresponds to

$$\mathcal{F}(\Psi)(f) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(f - C)^2}{1/(\pi^2 B^2)}\right]. \tag{71}$$

In the frequency domain the filter maintains its Gaussian shape. The term C corresponds to the Gaussian center, B is related to the function width $\sigma = (\sqrt{2}\pi B)^{-1}$.

6 Indeed by integrating on a rectangular path on the $\mathbb{C}$ plane the vertical edges give no contribution. The horizontal path corresponds to the gaussian integral on a horizontal line. By pushing this values to the limit we obtain the integration on the x-axis.

6.6.2 *Scales choice*

The relation between the filter parameters (B, C) and the scale factor a can be inferred from the analysis of the shape of $\mathcal{F}(\Psi_a)(f)$

$$\mathcal{F}(\Psi_a)(f) = \frac{1}{\sqrt{2\pi}} \exp\left[-\left(\frac{af - C}{1/(\pi B)}\right)^2\right] = \frac{1}{\sqrt{2\pi}} \exp\left[-\left(\frac{f - C/a}{1/(\pi a B)}\right)^2\right].$$

Both the central frequency and the filter width are compressed of the factor a. We write the center and the standard variation of the Gaussian curve depending on the scale parameter as f_c^a and σ^a. We obtain the following equivalence which allows us to understand the scale effect on the filter

$$f_c^a = \frac{C}{a} \text{ then, for } a = 1 \rightarrow f_c^1 = f_s, \tag{72}$$

$$\sigma^a = \frac{1}{\sqrt{2}\pi a B} \text{ then, for } a = 1 \rightarrow \sigma^1 = \frac{1}{\sqrt{2}\pi B}. \tag{73}$$

In the implementation we fix the first entry of the a array to the value $a_0 = 2.1$. This corresponds to the central frequency $f_c^{(a_0)} = 476$ Hz. We impose a great overlap of two filters with the requirement in Eq. 74. Given a scale a_j, the central frequency for the larger scale a_{j+1} must satisfy the condition below

$$\exp\left[-\frac{\left(f_c^{a_{j+1}} - f_c^{a_j}\right)^2}{2\left(\sigma^{a_j}\right)^2}\right] = 0.95. \tag{74}$$

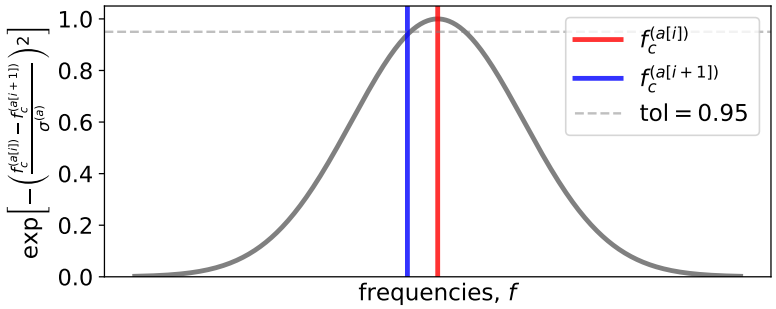

Figure 26: Given a fixed central frequency, we require to cover very tightly the spectrum, using the requirement that the central frequency at the next scale must be centered at a tolerance value, equal to 0.95.

In Figure 26 we plot the condition in Eq 74. The value of the central frequency at scale a_{j+1} must be centered in the point where the function value at scale a_j corresponds to 0.95.

By inverting the relation we find iteratively the central frequency values

$$-\frac{\left(f_c^{a_{j+1}} - f_c^{a_j}\right)^2}{2\left(\sigma^{a_j}\right)^2} = \ln(0.95) \rightarrow \left(f_c^{a_{j+1}} - f_c^{a_j}\right)^2 = -2\left(\sigma^{a_j}\right)^2 \ln(0.95)$$

$$f_c^{a_{j+1}} = f_c^{a_j} - \sigma^{a_j}\sqrt{-2\ln(0.95)},$$

we consider as solution the smaller frequency value as we increase the scale. Consequently, from Eq. 72 we extract the entries of the a array. The plots related to the central frequency and the spectrum width are shown in Figure 27. At

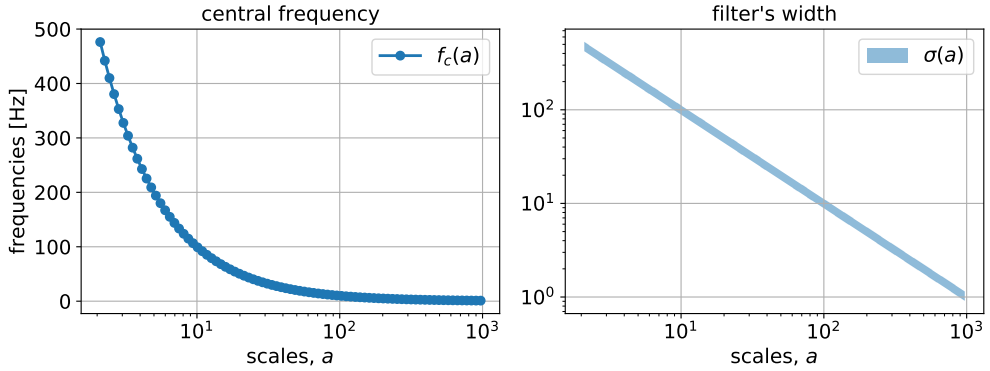

Figure 27: On the x-axis, index relative to the number of scales, $\#s = 83$, on the y-axis, frequency values. We cover the entire spectrum, from $\sim$ 470 Hz to 1 Hz. On the left we report the central frequency, on the right the correspondent filter width.

the highest scale, equivalent to $a = 970$, the correspondent central frequency is equivalent to 1.03 Hz, and the filter width to 0.23 Hz.

6.6.3 *Normalization based on Number of Contacts*

We add a normalization term on the loss function fixed as the number of training contacts for each patient $\#\mathcal{C}^{(p)}$.

$$f^{(p)}(S) = \alpha_0^{(p)} + \sum_{i \in \mathcal{C}^{(p)}} \left[\alpha_i^{(p)} \sum_{j=1}^{k} w_j K_j^{(p)} \left(S_i^{(p)}, S \right) \right] \tag{75}$$

$$\ell_{f^{(p)}} \left(S^{(p)}, y^{(p)} \right) = -\frac{1}{\#\mathcal{C}^{(p)}} \sum_{i \in \mathcal{C}^{(p)}} \log \left(1 + \exp \left(-y_i^{(p)} f^{(p)} \left(S_i^{(p)} \right) \right) \right) \tag{76}$$

The objective function assumes the following form

$$\begin{aligned} \underset{\alpha^{(1)},\ldots,\alpha^{(N)},w}{\text{minimize}} \Bigg\{ & \frac{1}{N} \sum_{p=1}^{N} \left(\frac{1}{\#\mathcal{C}^{(p)}} \left(\sum_{i \in \mathcal{C}^{(p)}} -\log \left[1 + \exp \left[-y_i^{(p)} f^{(p)} \left(S_i^{(p)} \right) \right] \right] \right) + \\ & \lambda \left(r_\lambda \|\alpha^{(p)}\|_1 + (1 - r_\lambda) \|\alpha^{(p)}\|_2^2 \right) \Bigg) \\ & + \beta \left(r_\beta \|w\|_1 + (1 - r_\beta) \|w\|_2^2 \right) \Bigg\} \\ & \text{s.t. } w_j \geq 0 \text{ for each } j = 1, \ldots, k. \end{aligned} \tag{77}$$

6.6.4 *Setup*

The implementation of CWT and the similarity matrices in Python speeded up the computational performance, thus allowing us to perform the experiments on the entire population of 59 focal epileptic patients. The library is available here `https://github.com/slipguru/mt-mkl`. We argue that the learning problem may get much harder given the heterogeneity of the population, in terms of class unbalance. We aim at observing if the method still provides an insight about the relevance of some frequency bands. The data representation and learning process is summarized in Figure 28.

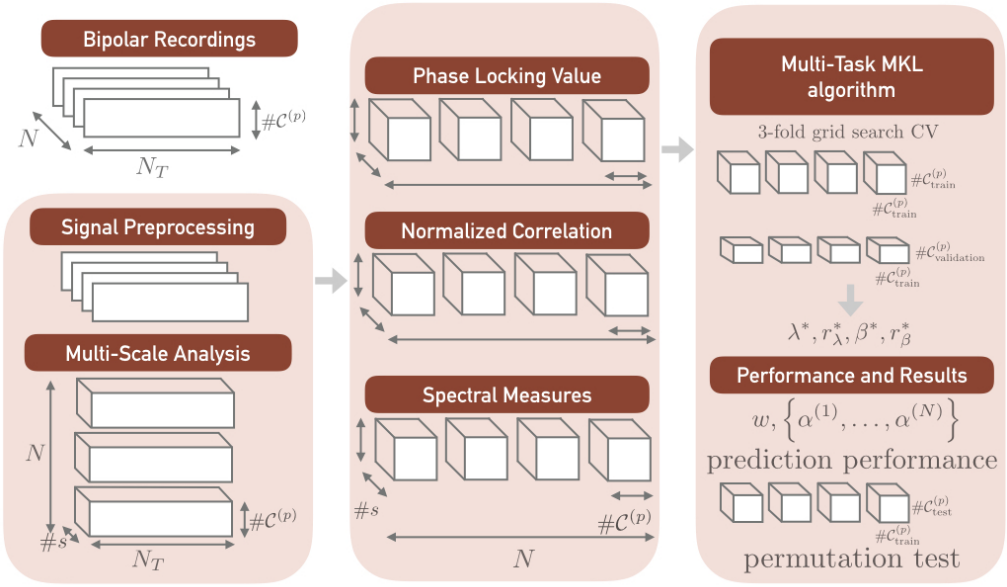

Figure 28: Schematic representation of the learning pipeline for Experiment II. From top left, the input data has dimension $N = 59$. In the middle column, we represent the data as in the previous experiment. We have in total $\#s \times 3$ similarity measures. In the last panel, the MT-MKL algorithm includes the minimization and the choice of the best model. This is tested on a test set. MT-MKL returns the set of kernels weights, the contacts weights, and the predictive result from the logistic probability function. We perform a permutation test.

Signal preprocessing is performed using the Python implementation of the Butterworth filter from the previous experiment [7]. The recordings have been reduced to a shared dimension on 590 seconds.

For what regards the multi-scale representation we resort to the array of scales derived from Formula 74. This leads to $\#s = 83$, number of wavelet

7 Last access October 19th, 2019. Butterworth filter `https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.butter.html`

scales. We built the similarity matrices as for the previous experiment. We construct the matrices related to the spectral measure by applying directly Formula 63.

Given the higher number of samples we are now able to split the dataset in learning and test sets and keep the latter for final evaluation of the results. We perform a stratified shuffled split which preserves the unbalance of epileptic and non epileptic contacts for each patient. During the learning procedure we determine the best model for each patient through a 3-fold grid search CV procedure. The hyper-parameters are selected in the following grid

$$\lambda = \beta = r_\lambda = r_\beta = [0.1, 0.4, 0.9]. \tag{78}$$

The selection takes place by considering the tuple with the highest averaged BA score of the 3-fold procedure.

$$(\lambda^*, \beta^*, r_\lambda^*, r_\beta^*) = \text{argmax}_{\lambda,\beta,r_\lambda,r_\beta} \left\{ \frac{1}{N} \sum_{p=1}^{N} (\text{BA score})_p \right\} \tag{79}$$

6.6.5 *Results*

We noticed a slow rate of convergence and for some runs of the experiment the model did not converge. If the minimization algorithm did not reach a minimum, within the tolerance value of 10^{-4}, it stops after a maximum amount of iterations, fixed at 200. To show the results we took into account the experiments where the balanced accuracy over the learning set exceeded the value 0.90. In Figure 29 we report the hyper-parameters selected across 16 repetitions of the experiment, from the condition in Eq. 79. We observe that the hyper-parameters related to the kernels weights (β^*, r_β^*) are stable across the repetitions.

6.6.6 *Lack of Selectivity*

We show in Figure 30 and in Figure 31 the frequencies selected respectively for the normalized correlation measure (in orange) and the phase locking value (in blue). We report the normalized weights related to bands selected across all the 16 repetitions of the experiment.

We expect the normalization to help us understanding if some of these rhythms are more interesting. To a higher value should correspond a higher relevance while in case of constant and equal weights across all features, all weights values should correspond to $1/(3\#s) = 0.004$. In this concern we do not observe the emergence of features selectivity in the discriminative task, for the entire set of 59 patients. The low selectivity of the algorithm can be here directly related to the selected regularization parameter r_β, which is the smallest among all the values in the grid (see Figure 29).

The predictive model relies over all the similarity measures. From Figure 30, across all the repetitions of the experiment, we observe a consistent pres-

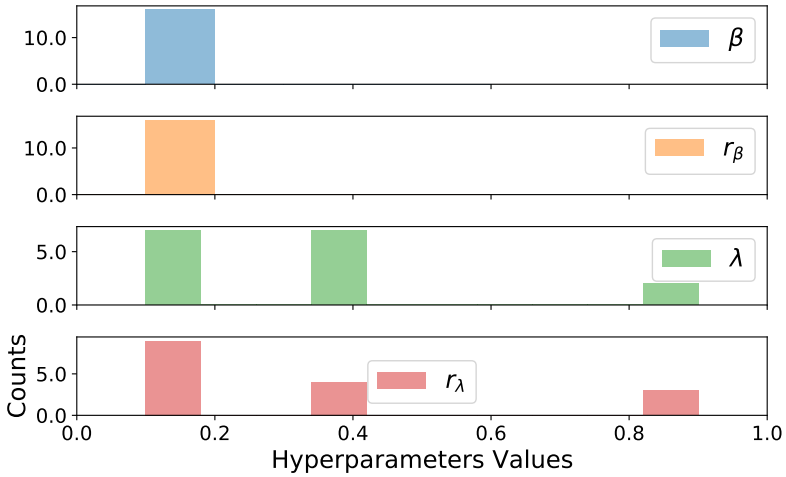

Figure 29: Histogram of best hyper-parameters selected across 16 repetitions of the experiment. The term r^*_β related to kernels selection is the lowest, indicating that the L^1 term in the elastic net has smaller relevance than the L^2. We observe a dependence from the dataset split for what concerns the selection of the tuple (λ^*, r^*_λ).

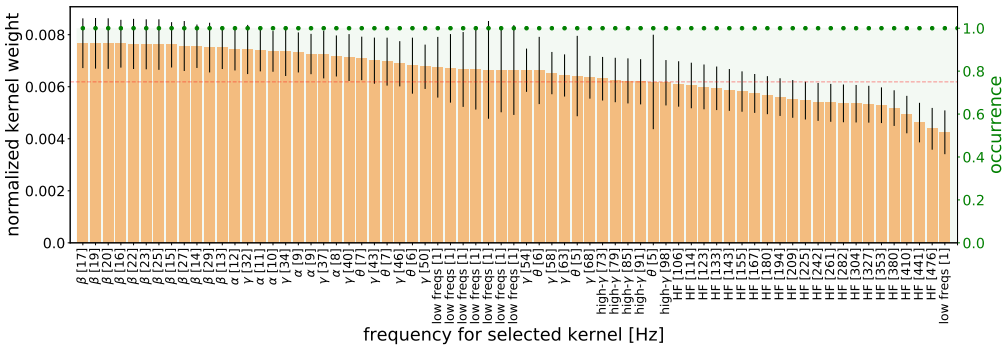

Figure 30: Normalized weights related to normalized correlation, selected at every repetition of the experiment. There is no selectivity but across all the repetition we observe a strong prevalence of β, γ rhythms, and high frequencies.

ence of β, γ, and high frequencies. Slower rhythms are selected across all the repetitions in Figure 31.

6.6.7 *Predictive Performance*

We evaluated the classification performance of the model. We computed the balanced accuracy score for each patient and then we average these values across all the population. The mean and the standard deviation related to this value, across the 16 repetitions of the experiment, are

$$\left\langle \frac{1}{N} \sum_{p=1}^{N} (\text{BA score})_p \right\rangle = 0.82 \pm 0.04 \tag{80}$$

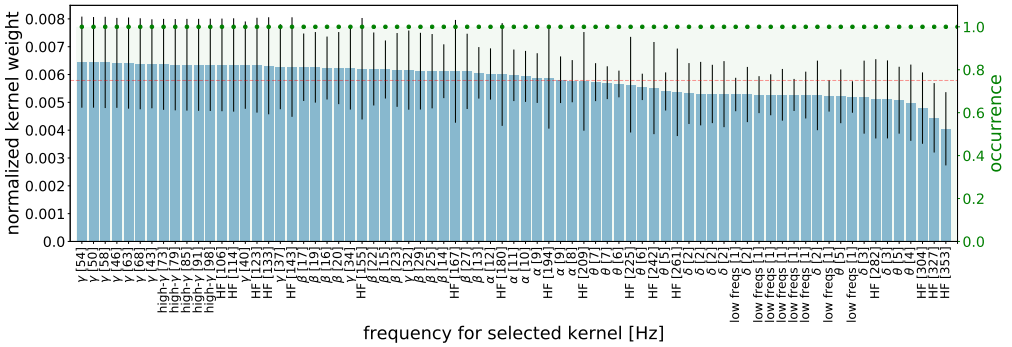

Figure 31: Normalized weights related to phase correlation, selected at every repetition of the experiment. We observe the presence of almost all the spectrum, from slow rhythms (δ) to high frequencies.

In Figure 32 we show other evaluation metrics, which allow to understand the predictive capacity of the model.

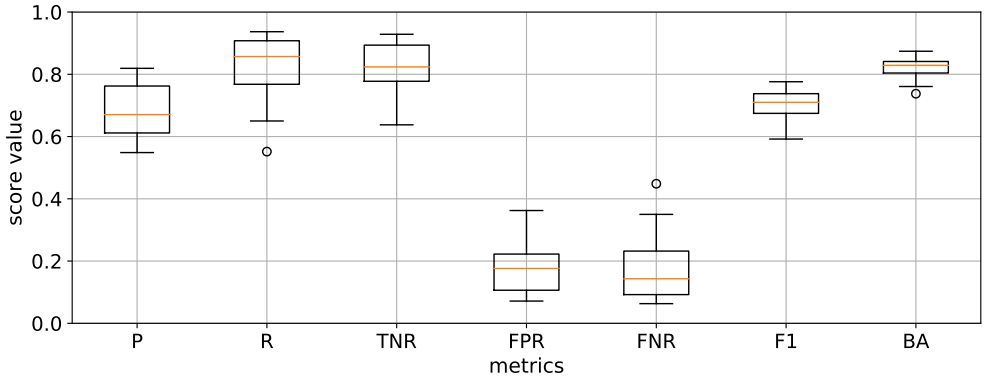

Figure 32: Performance evaluated on the test set across the 16 repetitions of the model, for which we fit the validation set. We observe that the metrics are all higher than chance.

P	R	TNR	FPR	FNR	F1	BA
0.68 ± 0.08	0.82 ± 0.11	0.82 ± 0.08	0.18 ± 0.08	0.18 ± 0.11	0.70 ± 0.05	0.82 ± 0.04

Table 22: Mean and standard deviation (round parenthesis) of several metrics evaluated on the test set. Starting from the left: precision, recall, true negative rate, false positive rate, false negative rate, F1 score, balanced accuracy.

In order to further assess the goodness of the prediction ability we perform a permutation test. This consists in the comparison between the prediction performance of the 16 repetitions of the experiment, and the prediction performance from another experiment, in the same setting, with randomly permuted y labels. We call *regular model* the ones obtained from the experiment in the standard setting, and *permuted model* the outcome of the learning process for the permuted labels. A high difference between the performance in the

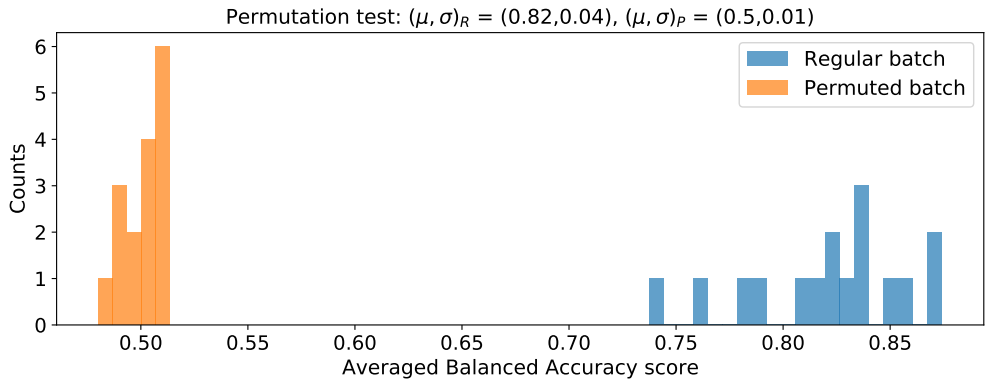

Figure 33: Result of the permutation test for the 16 selected models. By analyzing the prediction capacity of those models over the test set we exclude that these are extracted from the same distribution related to the permuted batch.

two settings is desirable. To quantify this we perform a 2-sample Kolmogorov Smirnov non-parametric test. The null hypothesis corresponds to balanced accuracy scores for the two models drawn from the same distribution. In Figure 33 we show distribution of the averaged across population balanced accuracy scores for the repetitions of the experiment. The orange bars correspond to the performance of the *permuted model*, in blue bars report the distribution of the *regular model* averaged scores. Given the small p-value $< 10^{-3}$ we discard the null hypothesis.

6.6.8 *Observations*

The analysis of the entire set of patients $N = 59$ gives negative result in terms selective capacity of the model. As we observed, this could follow from a higher variability of the class unbalance across the entire population. We observe nonetheless that the model, across repetitions of the experiment, maintains good predictive performance.

This result seems to corroborate the hypothesis that, in order to solve the predictive task, the method leverages on pathological information as well as spurious correlations. The latter do not capture the main characteristics of epilepsy, but help in solving the predictive task. The presence of spurious correlations is indeed favored by the form of the functional itself, which resorts to some of the bipolar recordings from a patient to assess the ones left as test.

It is hard to test this hypothesis and to quantify how spurious correlations contribute to the learning task. Reducing the size of the learning set would necessarily lead to a deterioration of the performance, whether the hypothesis holds true or not.

6.7 Experiment III: Labels Permutation and Role of Physiological Activity

In this part we address the limitations described in Section 6.5.3.

The main hypothesis here is that high predictive performance cannot prove the capacity of the model to capture features related to pathological activity, but rather its dependence on spurious correlations. Indeed the low selectivity in Experiment II may result from a wider contribution of the frequency spectrum but it may also denote the presence of spurious correlations. In this last part we examine the capacity of the model to rely on purely spurious correlations to achieve high performance.

To this aim, we randomize the labels for every contact and any patient.

After this operation, we expect any model trained on random labels to lack any features selectivity. We want to analyze nonetheless if the model performance is comparable to one without random permutation.

To this aim we compare two models, one trained with *random labels* and one with *permuted labels*. Given the two datasets, in this paradigm we split train, validation, and test set in the temporal domain.

For each patient, the similarity will be computed across the different chunks, by comparing the activity recorded from the same site at different time. We measure the capacity of the model to capture the input-output relation, in the regular case, and in the random case. If, despite the use of regularized methods, the model is complex enough to fit the dataset with random labels, we cannot discard the hypothesis that spurious activity may play a crucial role in the classification task.

6.7.1 *Setup*

We propose a scheme of the procedure in Figure 34. Starting from the left, the signal processing and multi-scale analysis is equivalent to the previous experiments. In the random label case, we permute the labels at the beginning of the process, before any preprocessing operation.

In the middle, for each bipolar recording we split the wavelet representation of dimensions $(\#s, N_T)$ in three chunks of equivalent length $(\#s, \text{int}(N_T/3))$. We denote these three periods as T_1, or train chunk, T_2, or validation chunk, and T_3, or test chunk. Given the length of the time recordings and the temporal resolution of the wavelet filter, there is not contaminations across different the three windows deriving from the convolution from the entire time series. At the lowest central frequency, equivalent to 1.03 Hz, the filter width equals 0.23 Hz.

We build the similarity matrices. In this experiment, all the matrices have dimensions $\#\mathcal{C}^{(p)} \times \#\mathcal{C}^{(p)}$, independently from the data split. We generate the similarity measures of the training step by comparing the activities across all contacts in the same window of time T_1. The similarity measures for the validation result from the comparison of the recordings in the period T_1 with

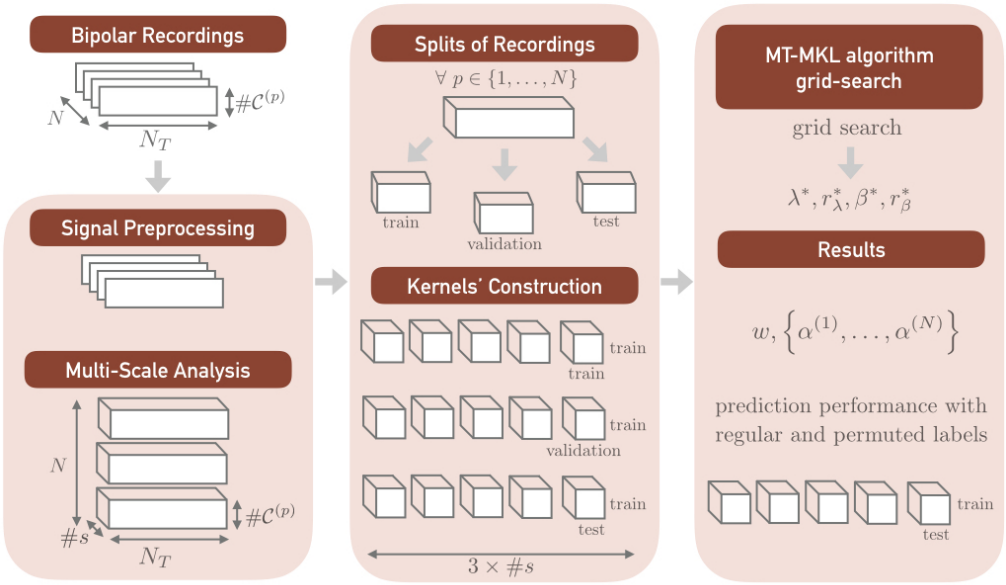

Figure 34: Experiment III: the preprocessing and the multi-scale representation is equivalent to previous experiments. The generation of the train, validation and test sets is such that we split the CWT representations in three blocks with the same length. Leakage effects to the split are negligible, given the filter resolution compared to the chunks length ($\sim$ 200 s). The kernel construction is such that we compare different time instants.

the ones at time T_2, while the test matrices derive from the comparison of the activity at time T_1 with the activity at time T_3.

We optimize the model by setting the grid of hyper-parameters to

$$\lambda = \beta = r_\lambda = r_\beta = [0.1, 0.4, 0.9]. \tag{81}$$

For each hyper-parameters tuple we do not perform a CV procedure, given the experiment design. The best tuple $(\lambda^*, r_\lambda^*, \beta^*, r_\beta^*)$ is chosen as the one which maximize the averaged balanced accuracy score. The maximum number of iterations for this experiment is fixed at 400.

6.7.2 *Results from Permutation*

We compare the classification results from the experiment with regular tags with the one with permuted tags. In the last case, three chunks from the same contact will share the same random tag, assigned before preprocessing. Due to the impossibility of repeating the data preparation step multiple times we did not perform the Kolmogorov Smirnov 2-sample. We perform the experiment by considering the results only if the value of balanced accuracy score

over the validation set was higher than 0.90 on the *regular model*. We observe already from two repetitions of the experiment (with different initialization values) that the performance of the *regular model* and the *permuted model* was equivalent. The averaged balanced accuracy scores over all patients in both cases exceed 90%.

BA score across patients	
regular	permuted
0.92 ± 0.13	0.95 ± 0.11
0.98 ± 0.06	0.95 ± 0.13

Table 23: Experiment III: classification performance on permuted and regular experiments. We observe comparable performance in both scenarios. The model is complex enough to fit a random relation and make prediction with optimal performance.

We cannot discard the possibility of the model to learn spurious correlation, despite the regularized approach. This result highlights the possible presence of spurious correlations among training, validation and test sets. Even though the hypothesis of spatial correlations is not confirmed we have the further evidence of strong temporal correlation, which enforces our worries about the role of correlation in this type of analysis.

6.8 Conclusions and Further Questions

In this Chapter we describe MT-MKL, our attempt to integrate information deriving from a multi-scale representation of the bipolar recordings, in order to classify pathological activity. The method has interpretable clinical outcome, regarding the selection of *similarity measures*, in phase and amplitude, and *frequencies*, which are the most relevant to the discriminative task. The first implementation of the method was run on a subset of patients with at least 25% of recordings belonging to the epileptogenic class.

From further investigations and a speed-up of the method performance, given by the Python implementation of the wavelet transform, we ran the algorithm on the entire population of 59 focal epileptic patients. Here, the unbalance between the two classes changes more abruptly than in the previous experiment and heavily impacts on the results, for what regards feature selection. The method does not show the emergence of the high frequencies only ($f > 100$ Hz), but we observe the selection of the entire feature set.

Despite the good classification performance and the distance to a random model in terms of predictive performance, we cannot discard the hypothesis of spurious correlation. On one hand this aspect does not affects negatively the predictive capacity of the method, which, leveraging on similarity, assigned labels with good classification performance within the same patient. On the other hand, this result leads to a loss of interpretability, as from a neurophysi-

ological perspective we cannot assess the reliability of feature selection for the characterization of the epileptic activity.

This consideration led us to the last experiment. Here, we split the recordings in three different chunks and we aimed at quantifying the predictive performance. Experiment III is, in our opinion, conclusive. The equivalent classification performance obtained from both the random and the regular batch shows that, despite the regularization framework, the method leverages on pure similarity at different frequency bands to learn the input-output relation, even when purely random.

We are left with several questions which we wish to address in the next chapters. Given that quantifying automatically the contribution of spurious correlation and physiological activity to the learning task was not possible, we first aim at finding a statistically reliable way to split the contributions from epileptogenic and physiological activity. Second, we wonder if using neurophysiological rhythms is the best way to proceed in the interpretation of pathological activity, or if those could guide to biased results. We tackle this question in Chapter 7. Third, we ask if MT-MKL, given its generality, can still be considered a useful method for feature extraction.

7

Search of Relevant Features and Stratification based on Post-Surgical Outcome

In this Chapter we answer some of the open questions from the previous analysis. As we have seen, contributions from physiological activity and the presence of spurious spatial correlations may potentially impact on the outcome, leading to biased and misinterpreted results. In the following we address these issues, resorting to a similar approach to the one adopted in Chapter 5. We improve the analysis by implementing a feature extraction method to quantify the effectiveness of our representation. In particular, we determine if an integral analysis of the bipolar recordings in the interictal phase may be as predictive as focusing on the presence of short bursts of high amplitude at different frequency bands. The analysis at the end of this chapter also assumes a clinical relevance, as in this last part we have been provided with the post-surgical outcome for a consistent portion of the population. We finally repeat the analysis on 25 *Engel I patients only.*

So far we defined data-driven models guided by few priors, without taking into account signal propagation. These models may nonetheless suffer of over-adaptation, and the discrimination of confounding factors from the true epileptic signal becomes extremely challenging. We identified several factors that represent obstacles to the analysis: (i) patterns variability, (ii) different pharmacological conditions, and (iii) focus position. In our limited knowledge, these aspects are linked. The patterns variability can indeed depend from the relative position between the contact and the neural population, the current direction and the volume through which the signal propagates (iii). Moreover point (ii) could affect (i): we expect mechanisms related to neurotransmission to be strongly modified by the AntiEpileptogenic treatment administered at the time of the acquisition of the signal.

As we do not take into account all these aspects, to minimize the complexity of our approach, we make in this Chapter a step back, in favor of linear approaches, where the input data is a preprocessed representation of the

recordings. If this implies on one hand less flexible models, on the other can potentially lead to higher interpretability and more reliable results.

To implement the following approach, the use of preprocessing and signal analysis techniques is key. In this regard, a short note about the importance of preprocessing is necessary. Preprocessing may indeed hide subtle pitfalls, as highlighted in [132], leading to catastrophic consequences, in term of biased and misinterpreted results.

In this work, several aspects of filtering, which previously we have not taken into account, are discussed. Among these, the use of Infinite Impulse Response (IIR) filters which should be avoided in favor of Finite Impulse Response (FIR) filters. The former show dramatic phase distortions if compared to the latter. Also the double filtering option used in the previous Chapter, which should prevent from these effects, leads to worst amplitude distortions than FIR filters. Not less important, bias may be also a consequence of difference implementations of the same filters, which constitutes a non-trivial issue in terms of reproducibility.

The goal for the following set of experiments is two fold: (i) understanding if some features may have higher relevance in the classification task, (ii) performing the analysis using an unbiased pipeline. For what concerns the former aspect, we divide the features in five subgroups. Some of these features represent an average measure of the neural activity, other quantifies the amount of time spent above threshold and are a measure of the abnormalities in the signal. The analysis will reveal which are more predictive for the EZ detection. For what regards (ii) we will not split the time series into smaller segments. We will compare the case in which the data split procedure suffers of contamination (different recordings acquired on the same subjects are both in the training as in the other splits) to the case more clinically plausible, where training, validation, and test sets are three different population subsets.

7.1 Extraction of Interpretable Features and Learning Pipeline

Guided by these considerations, for this analysis we rely on MNE[1], a standard, well established open-source Python library, developed specifically for the analysis of neurological datasets.

7.1.1 *Preprocessing and Feature Engineering*

For each bipolar recording we remove the contribution of slow rhythms and drifts using a FIR high-pass filter[2], with Hamming window, cut-off frequency

1 Last access: October 20th, 2019. MNE-Python library: `https://martinos.org/mne/stable/index.html`

2 Last access: October 20th, 2019. FIR on neural recordings `https://martinos.org/mne/dev/generated/mne.filter.filter_data.html`

1 Hz, transition bandwidth of 1 Hz. We apply a notch filter to remove the power line using a FIR notch filter with bandwidth equivalent to $f_{\text{line}}/200$, as in the standard MNE implementation[3].

The feature extraction from the generic bipolar recording $S_j^{(p)}$ after preprocessing replicates the one proposed in Chapter 5. We list in the following the feature categories and we describe any change:

(i) *first moments of the time series (mom)*;

(ii) *relative energy (fft)*, where the bands used to divide the spectrum are as in Table 15 and Table 16;

(iii) *normalized energy of wavelet coefficients and wavelet entropy (dwt)*, we resort on the PyWavelets [77];

(iv) *median absolute values (med)*, given a generic bipolar recording $S_j^{(p)}$ we estimate its baseline activity for each frequency band in Table 15 and 16. For each band we define a Hamming window FIR filter, whose shape depends on the band through the following relations

$$\begin{aligned}&\text{Left Bandwidth} = \min\big(\max(1/4 \cdot f_{\text{low}}, 2), f_{\text{low}}\big)\\ &\text{Right Bandwidth} = \min\big(\max(1/4 \cdot f_{\text{high}}, 2), f_{\text{Nyq}} - f_{\text{high}}\big).\end{aligned}$$

The Nyquist frequency, denoted as f_{Nyq} is equivalent to 500 Hz, f_{low} and f_{high} related to the cut-off frequencies, which are extreme values for the band Bk. Finally, we compute the median for the absolute value for the filtered bipolar recording at band Bk as in Eq. 82

$$\text{Abs Med}\left(S_j^{(p)}, \text{B}k\right) = \text{Median}\left|\text{FIR}_{\text{B}k}\left(S_j^{(p)}\right)\right|. \tag{82}$$

(v) *over-threshold activity (thr)*, it is an estimate abnormal activity for each bipolar recording. In particular we quantify the amount of time in which the bipolar recordings exceed a threshold at different frequency bands. As the physiological activity strongly varies across patients and brain regions, we quantify a baseline for each contact. In particular, given a bipolar recording $S_j^{(p)}$, we compute the threshold values Thr at a fixed band as

$$\text{Thr}\left(S_j^{(p)}, \text{B}k, c\right) = c \cdot \text{Abs Med}\left(S_j^{(p)}, \text{B}k\right), \tag{83}$$

with c integer in the interval $[3, 9]$. We extract each feature, defined as Time over Threshold (ToT) as

$$\text{ToT}\left(S_j^{(p)}, c, \text{B}k\right) = \sum_t \mathbb{I}\left[\left|\text{FIR}_{\text{B}k}\left(S_j^{(p)}(t)\right)\right| > \text{Thr}\left(S_j^{(p)}, \text{B}k, c\right)\right], \tag{84}$$

with $\mathbb{I}$ the indicator function.

3 Last access: October 20th, 2019. Notch filter `https://martinos.org/mne/dev/generated/mne.filter.notch_filter.html`

7.1.2 *Definition of Different Dataset Splits*

We compare two experimental setups rising from different dataset split strategies. The two scenarios are reported in Figure 35.

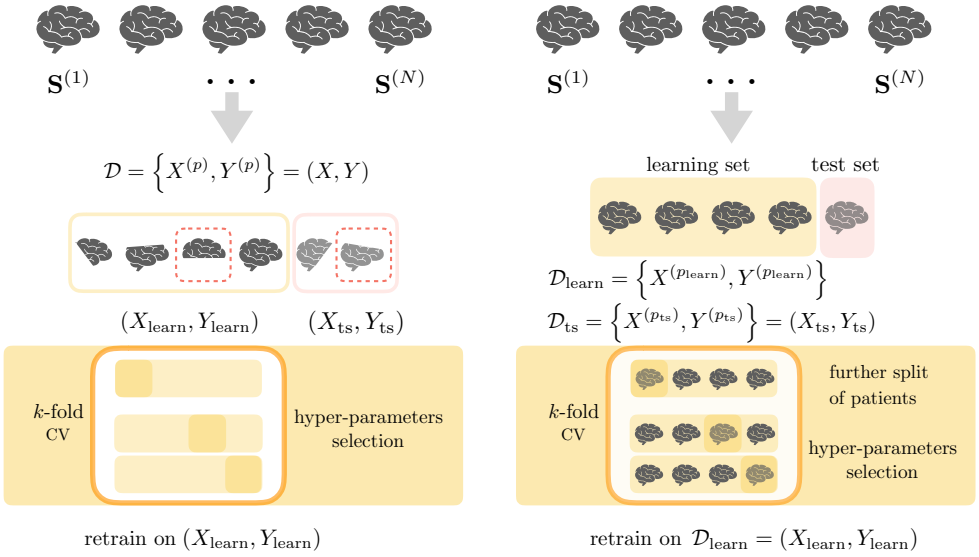

Figure 35: Bipolar recordings acquired from the same patient could potentially fall in the different splits, as shown with the dotted red box, on the left. Unbiased and clinical plausible scenario on the right. Here the population is split and recordings from test patients are separated from the learning set. The same strategy is applied on further internal splits.

On the left we apply the preprocessing and feature extraction steps on each bipolar recording. The tuple (X, Y) collects the representation for all the bipolar recordings with their corresponding labels. The dataset splits in learning and test sets, and further k-fold cross validation splits are performed without memory of the patient originating the samples at hand. This can cause contamination of the splits, as recordings from the same patients could potentially belong to both learning and test set. We call this scenario *recordings split*.

We depict in Figure 35, on the right, the unbiased case. Here we keep track of each patient before splitting in training, validation, and test set. We perform the preprocessing and feature extraction step, and we collect this output as a list of matrices, one for each patient in the population. We randomly sample a percentage of patients, which fall in the test set, and leave the others as data of the learning set. All the data forming the test set can be now put together in the tuple $(X_{\text{ts}}, Y_{\text{ts}})$. We perform further k-fold splits in the same fashion, by maintaining the separation of samples from different patients. We divide the learning set in k-folds and only at this point we generate the $(X_{\text{tr}}, Y_{\text{tr}})$ and $(X_{\text{vl}}, Y_{\text{vl}})$ tuples, by concatenating the features vectors from different subjects. We refer to this scenario as *patients split*. This last strategy assures the absence of biased results due to spurious spatial correlation, which could inflate the classification performance. To evaluate this possible effect, below we perform all the experiments by applying both dataset split strategies.

7.1.3 *Learning Strategies*

The dataset consists on $N = 59$ patients epileptic patients from Chapter 4. Given our representation, a generic feature vector from patient p and i-th recording is defined as $x_i^{(p)} \in \mathbb{R}^{148}$. We report the summary of our feature set in Table 24.

#features	*mom*	*fft*	*dwt*	*med*	*thr*
148	3	15	18	14	98

Table 24: Features subsets used during the analysis.

For simplicity in the next section we omit the superscript denoting the patient, as, once we generate the split of interest, we concatenate features coming from different patients.

7.1.3.1 *Experimental Setting and Metrics*

The experiments, unless when explicitly stated, will share the same setting. The different learning strategies presented here will leverage on the search of optimal hyper-parameters which is performed through a grid-search CV procedure. We report below more in detail how we operate the dataset split.

RECORDINGS SPLIT We split the recordings in learning and test sets, respectively 80% and 20% of the population. We select the best hyper-parameters through a stratified grid search 3-fold CV procedure. The dataset splits preserve the proportion of positive and negative samples of the original population.

PATIENTS SPLIT We randomly select 11 patients as test set ($\sim 20\%$) and the remaining 48 as learning set. In this setting we cannot take into account the unbalance between the two classes for each subset. We search the optimal parameter through 3-fold grid search cross validation, with 32 patients belonging to the training set and 16 to the validation set.

LEARNING METHODS AND HYPERPARAMETERS We use two standard methods to solve the binary classification task.

1. Logistic Regression with L^2 penalty, or LR-L^2

$$\min_{w,w_0} \left\{ \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log \left(\exp[-y_i (X_i^T w + w_0)] + 1 \right) \right\}. \tag{85}$$

2. Support Vector Machine with gaussian kernel, or SVM rbf

$$\begin{aligned}&\min_{w,b,\xi}\left\{\frac{1}{2}w^T w + C\sum_{i+1}^{n}\xi_i\right\}\\&\text{s.t. } y_i(w^T\phi(x_i)+b \geq 1-\xi_i)\\&\quad \xi_i \geq 0, \text{ for } i \in \{1,\ldots,n\}.\end{aligned}$$

For both classifiers the hyper-parameter C corresponds to the inverse of regularization strength. For what regards γ, variance of the Gaussian kernel in SVM, the value is automatically scaled based on the variance of the fitted dataset, as $\gamma = p \cdot \sigma(X)$. We choose the best hyper-parameter C^* among the ones in the C array that maximizes the mean value of the balanced accuracy score, using a 3-fold grid search CV procedure, where C is equivalent to

$$\begin{aligned}C =&[0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10, 20, 30, 40,\\&50, 60, 80, 100, 200, 300, 500, 750, 1000, 2000, 5000].\end{aligned}$$

EVALUATION METRICS We quantify the classifier performance through Precision (P), Recall (R), the Balanced Accuracy (BA), and F1 score (F1). We repeated the experiments 10 times to assess the results stability. Optimal hyperparameters are always chosen based on the maximum averaged balanced accuracy metric across the CV splits.

7.2 Experiment I: Prediction Performance using all Features

For this first experiment we resort to the entire feature set defined above, consisting on *mom*, *fft*, *dwt*, *med*, and *thr*.

We report the learning curves for one repetition of the experiment in Table 25. For each plot, on the x-axis, we report the hyper-parameter array C, on the y-axis, the averaged balanced accuracy score across the 3-fold repetitions. The yellow and blue curves show respectively the model performance on the training and validation set across the 3-fold cross validation procedure. The outcome of the CV procedure is the best C^* value, related to the highest mean balanced accuracy score. This is shown with a brown vertical line. In each row we present the result of a classifier, where we differentiate the dataset split. SVM rbf shows an increasing discrepancy between the training and the validation curves as we reduce the effect of regularization $(1/C)$. Logistic Regression with L^2 penalty is almost independent from the value of C.

In Table 26 we report the results obtained across 10 repetitions of the experiments. In the grey column we show the averaged balanced accuracy score of the validation set, from the 3-fold procedure. The other metrics are evaluated on the test set. In the upper part we show the results obtained for the *recordings split* setting, the lower part is relative to *patients split* settings. We observe

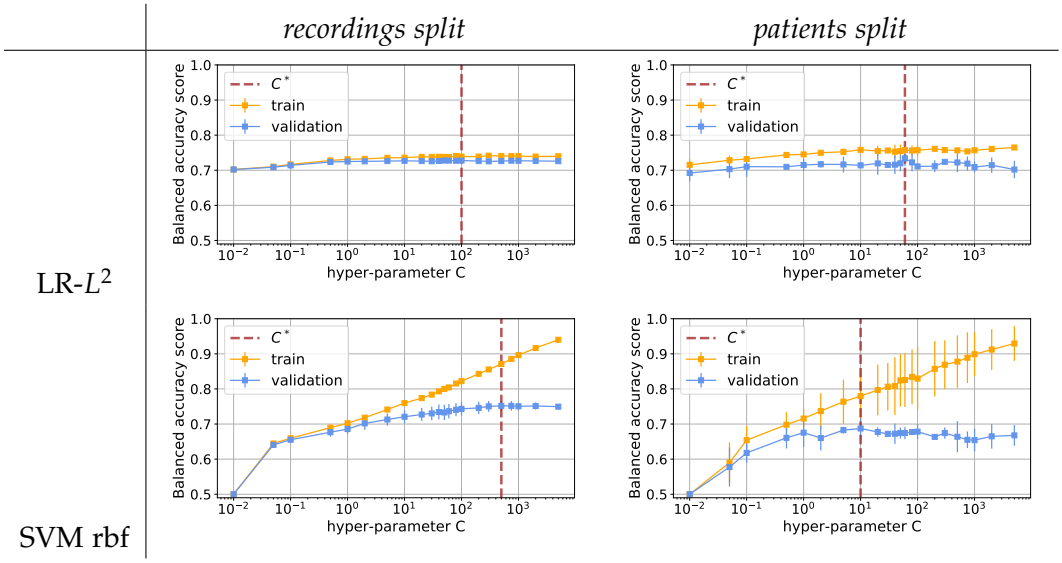

Table 25: Learning curves for one repetition of the experiment. On the x-axis, hyper-parameter C, on the y-axis, averaged balanced accuracy score across the 3-fold repetitions. On the left side we report the results related to recordings split, on the right, the results of the patients split procedure. The variances are higher in this latter case, but the balanced accuracy values are not dramatically different. The cause of higher fluctuations may originate from higher dependence on the split, given the absence stratification of the patients split scenario. The plots in the top row report the curves for the LR-L^2 models, while in the bottom row we report curves for the SVM rbf models, where σ depends on the fitted data. The higher discrepancy between the training and validation curves for SVM rbf show that this model is more prone to overfit than LR-L^2.

small variation between the performance for the two scenarios. The potential contribution of spurious correlation does not affect the results for this case.

recordings split					
model	$(\text{BA})_{\text{vl}}$	BA	F1	P	R
LR-L^2	0.722(4)	0.72(1)	0.61(2)	0.76(2)	0.50(1)
SVM rbf	0.754(4)	0.77(1)	0.69(2)	0.79(3)	0.61(2)
patients split					
model	$(\text{BA})_{\text{vl}}$	BA	F1	P	R
LR-L^2	0.714(8)	0.70(3)	0.56(7)	0.76(7)	0.45(8)
SVM rbf	0.700(6)	0.70(2)	0.56(4)	0.76(7)	0.45(5)

Table 26: Metrics scores for the two models. Upper rows, recordings splits, bottom rows, patients split. In both cases we report the balanced accuracy score correspondent to the optimal hyper-parameter, in the gray column. The other metrics are evaluated on the test set.

We observe nonetheless that the *recordings split* setting leads to more stable results, and slighly higher classification performance. In this scenario SVM rbf exceeds LR-L^2 in performance. In the *split patients* scenario, the performance of the classifiers are comparable and remain highly above chance level. The *recordings split* protocol, even if it does not separate recordings from the same patients across split, does not show inflated performance for the LR-L^2 model. The SVM is flexible enough to show slightly inflated results for the *recording splits* experiments, if compared to the classification performance of the SVM rbf, *patients splits* model.

7.3 Experiment II: Predictive Capacity of Features Subsets

The goal of this experiment is to apply a hard feature selection. We solve a different classification problem for each of the five subgroups of features in Table 24. In particular we apply the two classification methods presented in Experiment I, with identical C array and learning setup.

Here the dataset dimensionality changes depending on the considered features subset. Again, we compare the *recordings split* and *patients split* learning protocols.

Table 27 is the result of the learning procedure. We report performance of the classifiers across the ten repetitions of the experiment.

The table divides in two: the upper part reports the results of the *recordings split* protocol, while in the bottom we show classification performance obtained from the *patients split* learning protocol.

1. The performance of models related to the feature subgroup *thr* are comparable to the ones shown in Experiment I, across all metrics.

2. For models trained on *thr* features, SVM rbf shows higher classification performance in the *recording split* than in the *patients split* setting. This is not the case of LR-L^2, with exception of the precision score.

3. As in Experiment I, in the *patients split* protocol the standard deviations of all metrics, across the 10 repetitions, are much higher than in the other scenario.

This result leads us to further considerations about epileptic signatures. Pathological behavior of the neurological recordings should be searched among events exceeding baseline, without considering the average behavior at different frequency bands.

7.4 Experiment III: Automatic Feature Selection

As the information mostly resides in high amplitude patterns we focus here on the analysis of over threshold activity *thr* and discard the other features

recordings split						
subset	model	$(\text{BA})_{\text{vl}}$	BA	F1	P	R
thr	LR-L^2	0.689(3)	**0.68(1)**	**0.53(2)**	**0.80(3)**	**0.41(2)**
	SVM rbf	0.749(4)	**0.75(1)**	**0.65(2)**	**0.76(2)**	**0.57(2)**
dwt	LR-L^2	0.584(3)	0.584(6)	0.33(2)	0.55(3)	0.24(1)
	SVM rbf	0.532(8)	0.531(6)	0.15(2)	0.55(8)	0.09(1)
fft	LR-L^2	0.584(4)	0.585(8)	0.34(2)	0.53(3)	0.25(2)
	SVM rbf	0.538(3)	0.536(7)	0.17(2)	0.58(5)	0.10(1)
mom	LR-L^2	0.547(2)	0.548(5)	0.21(1)	0.61(3)	0.13(1)
	SVM rbf	0.594(7)	0.587(8)	0.32(2)	0.64(2)	0.21(2)
med	LR-L^2	0.573(3)	0.572(7)	0.28(2)	0.65(3)	0.18(1)
	SVM rbf	0.601(4)	0.601(7)	0.37(2)	0.56(2)	0.28(2)
patients split						
subset	model	$(\text{BA})_{\text{vl}}$	BA	F1	P	R
thr	LR-L^2	0.689(8)	**0.67(2)**	**0.50(6)**	**0.71(10)**	**0.39(7)**
	SVM rbf	0.704(6)	**0.68(3)**	**0.52(7)**	**0.76(6)**	**0.41(8)**
dwt	LR-L^2	0.59(1)	0.56(3)	0.27(7)	0.53(12)	0.19(7)
	SVM rbf	0.55(2)	0.53(2)	0.15(9)	0.53(2)	0.10(7)
fft	LR-L^2	0.59(1)	0.58(3)	0.32(10)	0.55(11)	0.25(11)
	SVM rbf	0.556(9)	0.53(2)	0.14(7)	0.62(10)	0.08(5)
mom	LR-L^2	0.557(6)	0.55(1)	0.21(5)	0.56(12)	0.12(4)
	SVM rbf	0.59(1)	0.59(2)	0.32(5)	0.65(8)	0.21(4)
med	LR-L^2	0.577(6)	0.572(7)	0.28(2)	0.65(3)	0.18(1)
	SVM rbf	0.59(1)	0.59(3)	0.35(7)	0.50(7)	0.28(6)

Table 27: Metrics scores for the two models learned on subsets of features. The upper rows refer to splits with mixed recordings, the bottom rows are related to splits with separated patients recordings.

subgroups. In particular, we leverage on an automatic feature selection approach which captures the most relevant features. We resort to a sparsity prior based on L^1 and L^2 penalties [138]. The combination of the two regularization terms is indeed convenient in presence of correlated variables. In our opinion this hypothesis holds for the *thr* features, as correlation may originate from: (i) their definition, at a fixed frequency band we expect that the time corresponding to the exceed of the threshold will be related for different threshold values; (ii) filters design, given mild transition bands there can be contributions from adjacent bands; (iii) the signal itself, if a periodic pattern consists in the sum of several sinusoidal oscillations, we expect to observe its contribution at different bands and similar effect is expected for non periodic patterns.

7.4.1 *Feature Selection Strategy*

Due to the difficulty in tuning the sparsity constraint in the standard implementation of the Elastic-Net on LR presented below

$$\text{argmin}_{w,w_0} \left\{ \frac{1-\rho}{2} w^T w + \rho \|w\|_1 + C \sum_{i=1}^{n} \log\left(\exp[-y_i(X_i^T w + w_0)] + 1\right) \right\}, \tag{86}$$

we decided to implement a nested feature selection method as proposed in the work of de Mol et al. [34].

$$\text{argmin}_{w} \left\{ \sum_{i=1}^{n} \mathcal{L}(w; y_i, X_i) + \mu \|w\|_2^2 + \tau \|w\|_1 \right\}, \tag{87}$$

with $\mathcal{L}$ loss function of a regression problem. The τ term imposes sparsity, while μ weights the contribution of the L^2 penalty.

The relation between the functional in Eq. 86 and the one in Eq. 87 can be obtained as

$$\text{argmin}_{w} \left\{ \frac{1-\rho}{2C} w^T w + \frac{\rho}{C} \|w\|_1 + \sum_{i=1}^{n} \log\left(\exp[-y_i X_i^T w] + 1\right) \right\}, \tag{88}$$

where we have the system $\tau = \frac{\rho}{C}$ and $\mu = \frac{1-\rho}{2C}$. The two optimization problems are equivalent for

$$\rho = \frac{\tau}{2\mu + \tau} \text{ and } C = \frac{1}{2\mu + \tau}. \tag{89}$$

Decoupling ρ and C parameters in the tuple (μ, τ) allows a higher control of the sparsity prior deriving from the L^1 over the smooth prior given by the L^2 term. We will resort next of this flexibility to implement a feature selection method with a nested approach.

LEARNING METHOD AND HYPERPARAMETERS The feature selection is based on a nested approach, which consists of a two-stage procedure, relying on three hyper-parameters (μ, τ, λ). The main change to the original algorithm consists in the use of the logistic regression loss. The two-stage method is reported in Equations 90, 91, 92, and 93.

- *stage 1*: $\mu = \mu_0$ across the CV procedure, hyper-parameters (τ, λ) are selected using a 3-fold CV procedure

$$\arg\min_{w} \left\{ \sum_{i=1}^{M} \log\left(\exp[-y_i X_i^T w] + 1\right) + \mu_0 \|w\|_2^2 + \tau \|w\|_1 \right\} \tag{90}$$

$$\arg\min_{\tilde{w}} \left\{ \sum_{i=1}^{M} \log\left(\exp[-y_i \tilde{X}_i^T \tilde{w}] + 1\right) + \lambda \|\tilde{w}\|_2^2 \right\}. \tag{91}$$

We solve two minimization problems. In Eq. 90 we minimize the Elastic-Net problem where we enforce the sparsity constraint. This is obtained by fixing $\mu_0 \ll \tau$. By forcing selectivity, we obtain a subset of features. We denote $\tilde{w}$ as the array of weights with non null components and $\tilde{X}$ as the correspondent data matrix of reduced dimensionality in the feature space. In Eq. 91 we solve the classification problem in the smaller dimensional feature space, with the imposition of the L^2 prior. The outcome of this procedure is the optimal tuple (τ^*, λ^*) which maximizes the averaged balanced accuracy score over the validation set.

- *stage 2*: (τ^*, λ^*) fixed from stage 1, $\forall \mu_j > \mu_0$, with (X, y) entire learning set

$$\arg\min_{w} \Big\{ \sum_{i=1}^{M} \log\Big(\exp[-y_i X_i^T w] + 1\Big) + \mu_j \|w\|_2^2 + \tau^* \|w\|_1 \Big\} \tag{92}$$

$$\arg\min_{\tilde{w}} \Big\{ \sum_{i=1}^{M} \log\Big(\exp[-y_i \tilde{X}_i^T \tilde{w}] + 1\Big) + \lambda^* \|\tilde{w}\|_2^2 \Big\}. \tag{93}$$

  As defined, this approach leads to a set of models, of increasing dimensionality, with the relaxation of the sparsity constraint in favor of the L^2 term, see Eq. 92. The result of this operation is the selection of a subset of features. Here we use the notation $\tilde{w}$ to denote the array of weights corresponding to non null components. For each μ value we obtain a classifier, as result of Eq. 93.

As the value of μ increases the algorithm selects a larger subset of features. This leads to high interpretability of the results as features selected from the sparsest models should be the most relevant to the classification task. As we relax the L^1 constraint we expect the model to catch further correlated features to those selected in the sparse model. Moreover we expect in the asymptotic regime to obtain comparable predictive performance [34] [33]. The hyper-parameters arrays μ, τ, and λ have respectively 45 entries equally logarithmic spaced in the range $[10^{-5}, 3 \cdot 10^4]$, 10 entries equally logarithmic spaces values in the range $[1, 10^2]$, and 10 entries equally logarithmic spaced values in the range $[0.1, 10^2]$. The dataset splits are equivalent to the ones defined for Experiment I and Experiment II.

EVALUATION METRICS Again we compute the same scores as in Experiment I and Experiment II. The hyper-parameters selection should rely on the highest value of averaged balanced accuracy score on the CV splits. Nonetheless, we observe that the performance are almost equivalent across the different values of (τ, λ) with fluctuations which are smaller than $< 2\%$. For this reason, based on the validation set, we kept the sparsest model, correspondent to $\tau^* = 100$ and we select the best λ^*, correspondent to the best balanced accuracy score, across all the experiments. We ran the experiments 10 times, for both recordings split and patients split scenarios.

7.4.2 *Results*

In Figure 36 we report the classification performance of the procedure over the second stage across 10 repetition of the experiment. These results correspond to fixed τ^* and λ^* values, with the former always constant. We test the models as we increase the μ value, or the importance of the L^2 regularization term over the sparsity term. We observe that: (i) as from the previous experiments, the performance for the recordings split scenario are not significantly higher; (ii) as expected asymptotically, the performance is independent from the μ value, which guarantees model stability.

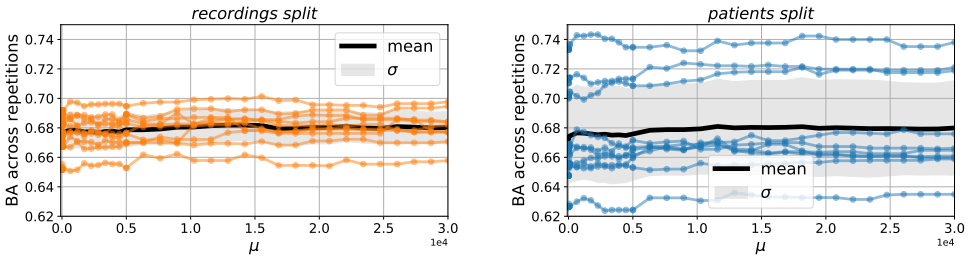

Figure 36: Predictive performance of the two-stage procedure, for the 45 models obtained for different values of μ. On the the x-axis, the hyper-parameter μ, on the y-axis, the balanced accuracy score. The colored curves are the results of each run of the experiment. The black curve and the gray area report respectively the mean and standard deviation of the balanced accuracy across ten repetitions. In general, we observe that the performance is not affected by the model size, which assures the stability of the results. On the left *recordings split* setting, on the right *patients split*. The *recordings split* shows lower variance.

We report the number of features as we increase the value of μ, across the 10 repetitions of the experiment in Figure 37. We observe no discrepancy between the two splitting data paradigms.

As the ratio between the L^1 and L^2 hyper-parameters decreases, the number of features selected by the algorithm increases. In this concern, we remind that $\tau^* = 100$ across all the experiments. For the smallest μ value, the number of selected features is the minimum across all the experiments. This represents the sparsest scenario. We call the features selected in this regime *ancestors*.

We show in Figure 38 the occurrence of the ancestors for models obtained in the *recording split* protocol, at the top, and *patients split* protocol, at the bottom. On x-axis we put the features, which are identified by the band Bk and the constant value c. We observe that across learning protocols and repetitions high threshold values c are never selected at high frequency bands ($f > 290$ Hz). The feature selection for *recordings split* and *patients split* is consistent. The most selected features correspond to the ones in the α, β, and γ rhythms.

We show in Table 39 the regularization paths for the most selected rhythms. This is the result for one repetition of experiment, from the *recordings split* protocol. The regularization path is informative to establish a hierarchy among features. On the x-axis we report the μ parameter, on the y-axis the coefficient

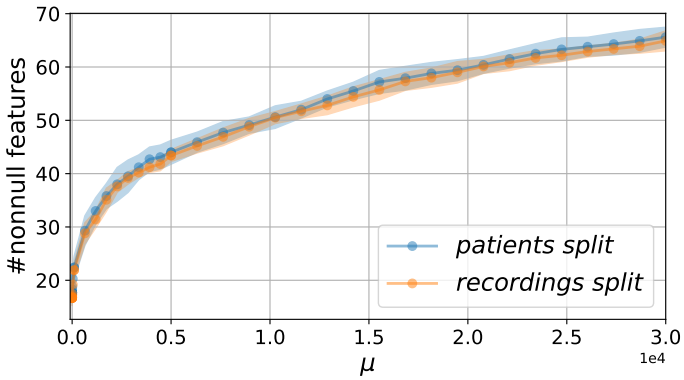

Figure 37: Number of selected features as function of μ. On x-axis the μ coefficient, on the y-axis the number of nonnull coefficients across repetitions of the experiment. The mean and standard deviation are respectively represented with dot markers and colored areas. We report the *recording split* and *patients split* respectively in orange and blue.

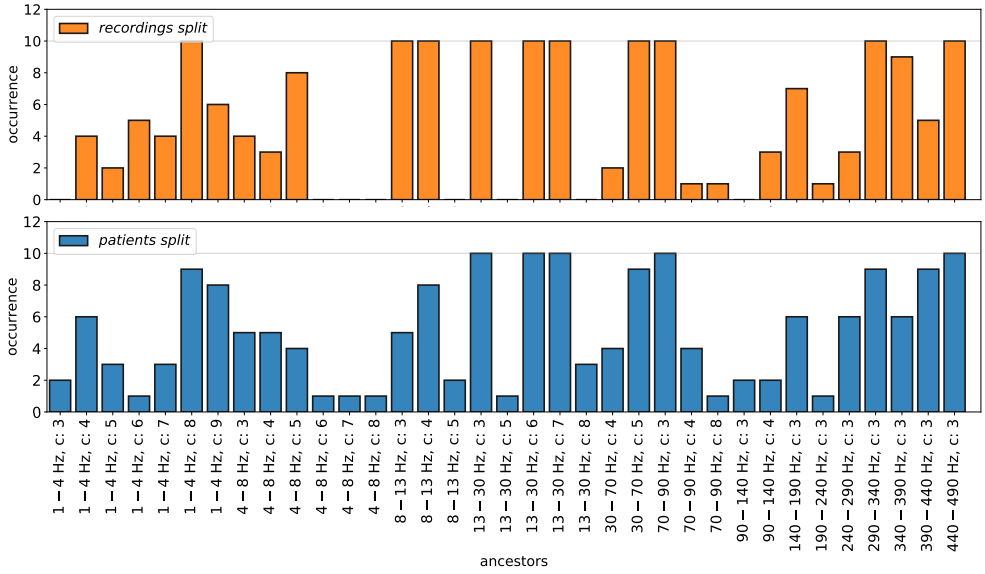

Figure 38: Histogram for the occurrence of the ancestors across ten repetitions of the experiment. Top: *recordings split*, bottom: *patients split*. The gray line denotes the maximum occurrence value. We observe that the last scenario leverages on a smaller amount of features, but there is a good agreement on the feature importance.

values. The ancestors have a regularization path which never goes to zero, even for the smallest μ value.

The ancestors selected in the same experiment are shown in Figure 28. Here we emphasize the correlation among these features, using a heatmap. On the left we put the *ancestors*. The ancestors are $1-4$ Hz $c=5$, $c=8$; $4-8$ Hz $c=3$, $c=5$; $8-13$ Hz $c=4$; $13-30$ Hz $c=3$, $c=6$, $c=7$, $c=8$; $30-70$ Hz $c=5$; $70-90$ Hz $c=4$; $90-140$ Hz $c=4$; $140-190$ Hz $c=3$; $290-340$ Hz $c=3$; $340-390$ Hz $c=3$; $390-440$ Hz $c=3$; $440-490$ Hz $c=3$. The

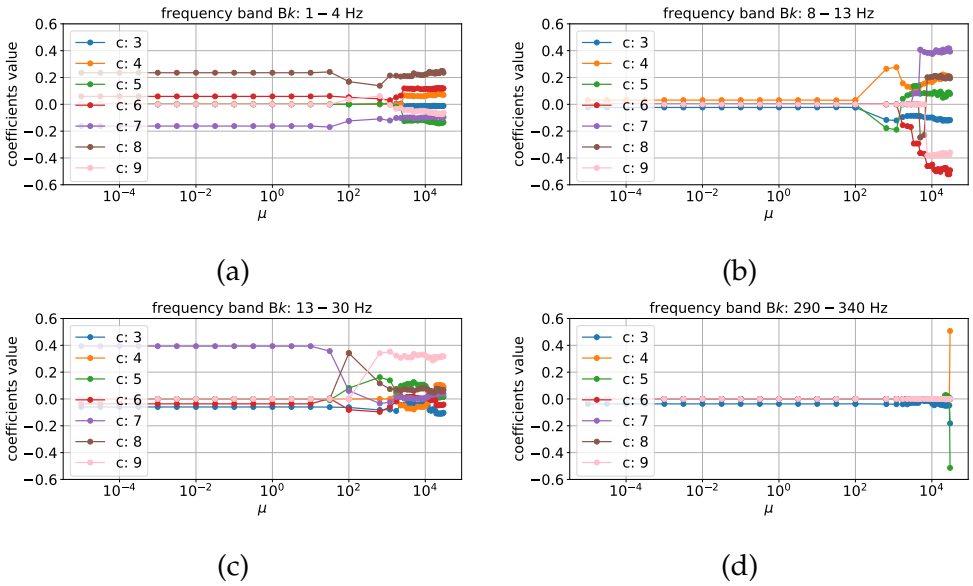

Figure 39: Regularization paths for some of the most recurrent bands, for one repetition of the experiments. (a) δ rhythm, (b) α rhythm, (c) β rhythm, and (d) $[290, 340]$ Hz.

black lines denote the division among blocks of coordinates, each black line is related to an ancestor. We then order the features of the largest model, based on their correlation with the ancestors, where the features are grouped to the ancestor for which we measure the highest correlation.

Despite almost all the frequency bands have an ancestor in the sparsest model, we observe that, by dividing the correlation matrix in the two classes, there is higher correlation among features for the positive class, especially in the high frequency bands ($f \in [70, 240]$ Hz).

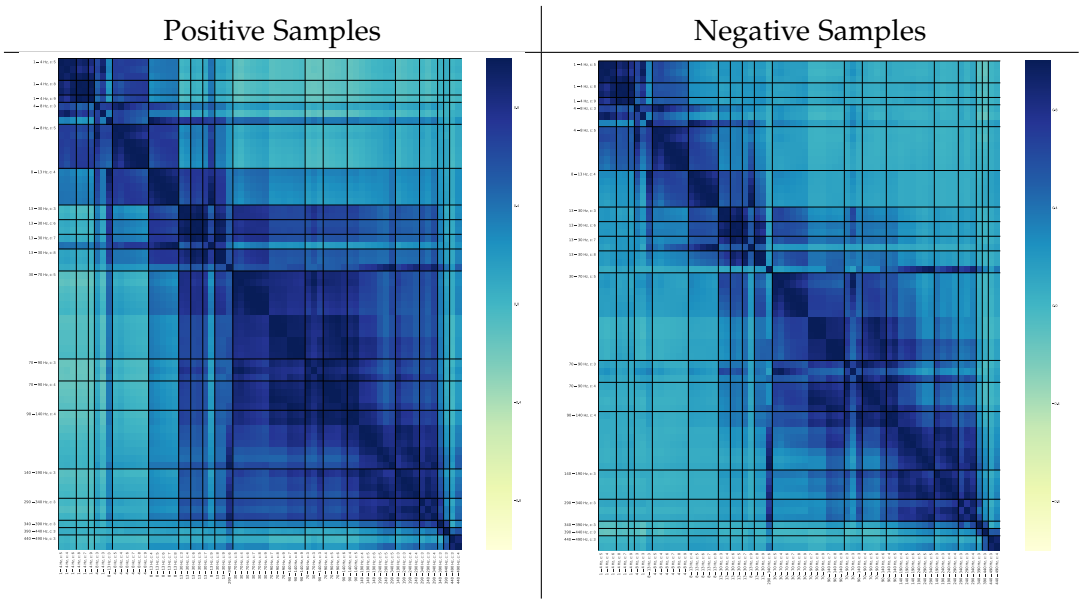

Table 28: Correlation matrices for the family of **thr** features selected through the one repetition of the experiments with nested features, for a split performed across patients, for the model with the biggest μ. On the left, the correlation is computed between epileptic contacts only, on the right, there are non epileptic contacts. The black lines denote the ancestors features, selected in the sparsest model ($\mu = 10^{-5}$). Those are in order: $1-4$ Hz $c = 5$, $c = 8$; $4-8$ Hz $c = 3$, $c = 5$; $8-13$ Hz $c = 4$; $13-30$ Hz $c = 3$, $c = 6$, $c = 7$, $c = 8$; $30-70$ Hz $c = 5$; $70-90$ Hz $c = 4$; $90-140$ Hz $c = 4$; $140-190$ Hz $c = 3$; $290-340$ Hz $c = 3$; $340-390$ Hz $c = 3$; $390-440$ Hz $c = 3$; $440-490$ Hz $c = 3$. We measured the correlation between each feature of the largest model and the ancestors, and we cluster the former depending on the highest correlation value. In this descriptive result, there are indeed some differences between the positive and the negative classes. The epileptic contacts show higher correlations between features related to high frequencies.

7.5 Clinically Unbiased Results: Engel I

All the previous experiments were based on pre-surgical assessment of the EZ, which we translated in binary labels as commented in Chapter 4.

In the last part of the thesis we have been provided with the post-surgical assessment for a substantial portion of the population. This information includes:

(i) the type of *surgical intervention*, which differentiates in ablation, thermocoagulation, or not operated;

(ii) the *removed region* in case of ablative surgery;

(iii) the *post-surgical outcome* in case of resective surgery or thermocoagulation, which consists in the Engel classification;

(iv) and the *anti-epileptogenic treatment* administered at the time of the acquisition.

Based on the post-surgical outcome the population stratifies as in Table 29.

	Engel I	Engel II	Engel III	Engel IV	Unknown
#patients	**25**	6	4	6	18

Table 29: Stratified population based on post-surgical outcome. Engel I patients, in bold, constitute the subgroup for which the pre-surgical evaluation has been effective. The category unknown includes both the patients who refused the surgery as the ones for which the post-surgical classification outcome is not available.

The post-surgical evaluations give higher relevance to the analysis, as we are able to consider only the subpopulation for which the pre-surgical assessment and the surgical intervention have been effective. With the goal of repeating the analysis presented in the first part of this Chapter, we first highlight some aspects: (i) positive contacts position and ablated regions approximately coincide for all patients belonging to the Engel I class. A direct correspondence between pre-surgical assessment and resected region is nonetheless not straightforward, because of the higher spatial resolution about the contact position of SEEG compared to the size of entire brain regions of ablated tissue. This approximation should not affect thermo-coagulated cases, but these are nonetheless extremely rare.The contacts position as the region for each contact have been extracted as for Chapter 5, using the Destrieux atlas[4]. (ii) Negative post-surgical outcomes do not necessarily correspond to miscorrect pre-surgical evaluations. We notice that, in some cases, the pre-surgical assessment on Engel IV patients defines as epileptogenic a large number of contacts falling in different regions, which have not been entirely ablated. This may be related to

4 Last access: November 24th, 2019. Cortical parcellation from Free Surfer. `https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation`

further pre-surgical evaluations about cognitive damages deriving from such drastic resections.

Given the impossibility of taking into account these further factors, we decide to exclude from the following analysis all the classes but Engel I. We limit the analysis to the *patients split* paradigm.

7.5.1 *Experiment I: Prediction Performance using all Features*

We repeat the analysis of Section 7.2, over the entire set of features defined before.

7.5.1.1 *Experimental Setup*

As in Section 7.2, we resort to LR-L^2 and SVM rbf classifiers. The hyper-parameter array C is identical to the one proposed in the previous version of the experiments. Similarly to the previous experiments, the split in learning and test sets respects the ratio 80%/20%. As the dataset consists of $N = 25$ patients, 20 patients belong to the learning set while 5 fall in the test set. During learning, we fix the optimal hyper-parameters through a 4-fold grid search CV procedure. If not specified, the optimal hyper-parameter is the one which maximizes the averaged balanced accuracy score across the 4 validation folds. The model is then retrained over the learning set and tested. We repeat each experiment ten times to assess the stability of the performance.

7.5.1.2 *Results*

patients split					
model	$(\text{BA})_{\text{vl}}$	BA	F1	P	R
LR-L^2	0.727(9)	0.72(6)	0.57(12)	0.65(10)	0.54(15)
SVM rbf	0.718(7)	0.68(4)	0.52(9)	0.78(7)	0.40(10)

Table 30: Experiment I, Engel I class. Scores for the two models. We evaluate the performance over 5 patients, across ten repetitions of the experiment. We separate patients across the learning and test procedures.

We observe high fluctuation of the metrics scores. This is probably due to the reduced amount of samples, if compared with the results in Table 26. The results do not significantly improve because of the stratification based on post-surgical outcome.

7.5.2 *Experiment II: Predictive Capacity of Features Subsets*

As before, here we evaluate if there is a more predictive subset of features. The experimental setup is identical to the one of Experiment I presented above.

7.5.2.1 *Results*

In Table 31 we show the predictive performance for the models. These results

		patients split				
subset	model	$(\text{BA})_{\text{vl}}$	BA	F1	P	R
thr	LR-L^2	**0.71(2)**	**0.68(5)**	**0.52(11)**	**0.76(10)**	**0.41(11)**
	SVM rbf	**0.72(2)**	**0.71(4)**	**0.58(8)**	**0.82(5)**	**0.46(8)**
dwt	LR-L^2	0.609(8)	0.58(5)	0.30(15)	0.64(15)	0.23(15)
	SVM rbf	0.59(1)	0.53(3)	0.15(10)	0.58(16)	0.09(7)
fft	LR-L^2	0.612(7)	0.57(5)	0.29(12)	0.56(14)	0.22(15)
	SVM rbf	0.62(1)	0.60(5)	0.36(14)	0.55(8)	0.31(17)
mom	LR-L^2	0.578(6)	0.55(3)	0.22(10)	0.57(18)	0.16(11)
	SVM rbf	0.56(2)	0.52(3)	0.08(10)	0.63(38)	0.05(7)
med	LR-L^2	0.61(1)	0.59(5)	0.33(12)	0.66(15)	0.26(17)
	SVM rbf	0.61(2)	0.58(4)	0.32(13)	0.59(14)	0.27(15)

Table 31: Experiment II, Engel I class. Scores for the two classification methods, trained on single features subgroups. We evaluate the performance on 5 patients, across 10 repetitions of the experiment. Again, we observe the emergence of the subgroup *thr* as the most relevant to the classification task.

must be compared to the experiment in Table 27. As before, we confirm the emergence of *thr* as the most relevant feature subgroup, among the proposed ones. By comparing the results from Experiment I with results for the *thr* features in Experiment II, we notice that the gap between balanced accuracy score evaluated on the validation test and on the test set is almost equivalent, revealing the stability of the result for this subgroup. Moreover we observe that the precision score increases for both the LR-L^2 and SVM rbf models. This corresponds to a decrease of the number of false positive samples. Statistical fluctuations are nonetheless too high to assess the significance of this result.

7.5.3 *Experiment III: Automatic Feature Selection*

As before, we aim here at finding frequency bands of higher relevance, resorting to the nested two-stage method. The learning protocol is identical to the one of Experiments I and II from this section. The hyper-parameter arrays are equivalent to the one in Experiment III, Section 7.4.

By design the choice of the optimal tuple (τ^*, λ^*) should be based on the highest averaged balanced accuracy score. As before, we notice small variations $< 2\%$ of the scores across the hyper-parameters grid. We then consider the sparsest model, constraining $\tau^* = \max(\tau)$. We determine the λ^* value as the one which maximizes the balanced accuracy score over the 4 validation folds.

7.5.3.1 *Results*

We observe from Figure 40 that the performance is stable across the entire set of μ parameters. This guarantees that the input-output relation is stable and

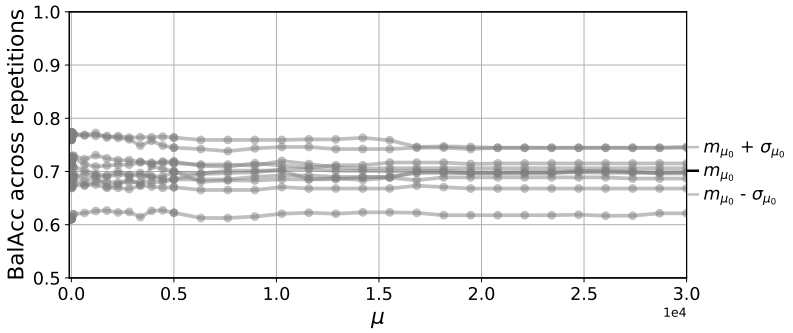

Figure 40: Balanced accuracy curve evaluated on the test set and μ increases for Experiment III, Engel I class. Each gray curve represents a single repetition of the experiment. The values m_{μ_0} and σ_{μ_0} represent respectively the mean value and the standard deviation of the balanced accuracy score for the sparsest model, reported in Table 32.

robust despite the increasing dimensionality. In Table 32 we report the metrics scores for the two-stage feature selection model. We observe a slight increase

patients split					
model	BA_{vl}	BA	F1 score	P	R
L^1L^2-LR	0.71(1)	0.70(4)	0.57(9)	0.83(8)	0.43(9)

Table 32: Experiment III, Engel I class. Scores evaluated on five test patients averaged across ten repetitions of the experiment for the models with $\mu = \mu_0$. We report the mean value and standard deviation (in parenthesis).

of the learning performance, if compared to the LR-L^2 model of Experiment II, first row in Table 31.

We observe stability also in the number of selected features across repetitions of the experiment. In Figure 41, we show the number of features as the μ value increases. As μ increases, we notice that curve related to nonnull features tends to a plateaux. This result is encouraging, as, despite the increase of the μ value, the models do not show a further increase of complexity.

We show in Figure 42 the histogram reporting the ancestors occurrence across the ten repetitions of the experiment. We observe a coherent use of features at the neurophysiological bands, especially for δ, β, and γ bands. The contribution deriving from very high frequencies ($f > 200$ Hz) is always present, across every repetition of the experiment.

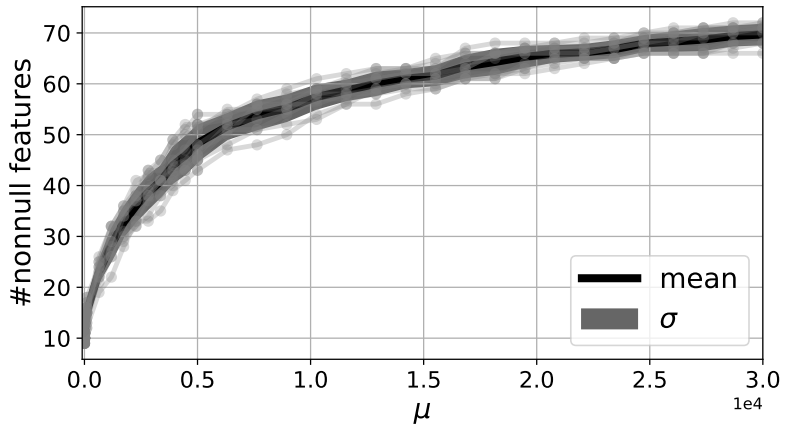

Figure 41: Experiment III, Engel I class. Number of selected features as the μ value increases. The black line and gray area denote respectively the mean and standard deviation of the number of selected features, the curves reports the result of each repetition.

7.5.4 *Experiment IV: Personalized-Models using Automatic Feature Selection*

As a final analysis we consider a patient at the time, to build personalized-models and to evaluate the stability of the feature selection approach across patients. We have here multiple goals: (i) confirming the importance of the *thr* set; (ii) observing if the classification performance is uniform or the difficulty of the classification task varies across patients; (iii) understanding if this last aspect is related to the unbalance of the classification problem.

7.5.4.1 *Experimental Setup*

Given a patient p, we split the dataset $(X^{(p)}, y^{(p)})$ in learning and test sets consisting respectively of 80% and 20% bipolar contacts, while maintaining the class unbalance. We are aware that this may lead to overestimation of the classification performance, as we cannot exclude contributions from spurious spatial correlations. The μ, τ, and λ arrays consist respectively of 15, 10, and 10 logarithmically spaced values in the intervals $[10^{-5}, 10^{3}]$, $[1, 100]$ and $[0.1, 100]$. We select the optimal $\left((\tau^{(p)})^*, (\lambda^{(p)})^*\right)$ with a 5-fold CV procedure. For each patient we repeat the experiment ten times.

7.5.4.2 *Results*

In Table 33 we report the classification performance obtained from this single model pipeline. The subject ID refers to the dataset presented in Chapter 4. Unbalance refers to the proportion of examples from the epileptic class over the total. $\langle \mathrm{BA}_{\mathrm{sparse}} \rangle$ and $\langle \mathrm{BA}_{\mathrm{large}} \rangle$ are respectively the balanced accuracy scores averaged over the ten repetitions of the experiments, respectively for the sparest model, with $\mu = 10^{-5}$ and the largest one, with $\mu = 10^{3}$. As in Experiment

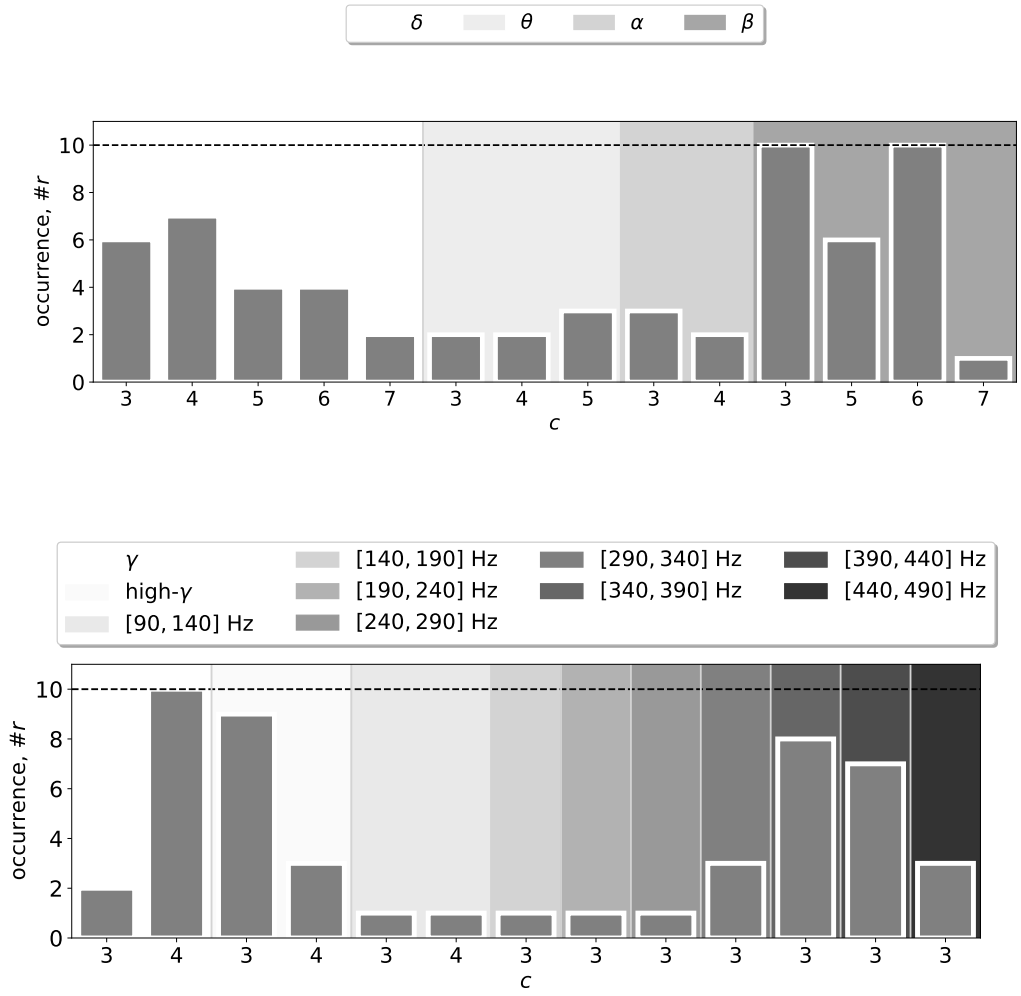

Figure 42: Experiment III, Engel I class. Occurrence of ancestors across ten repetitions of the experiment. The bar corresponds to the number of time the feature has been selected in the sparsest scenario. On the x-axis we put the c constant. This value, together with the legend, identifies a unique feature. The black dashed line corresponds to the maximum value of the histogram, equal to 10. We observe that over-threshold activity at β and γ bands is selected across all repetitions.

III, we notice that the balanced accuracy score in stable independently from the μ value. This result shows the stability of our models, even if the number of samples for this scenario is smaller than in the previous cases. We observe a strong variability of the models classification performance across patients. For some subjects the clinical features taken into account are discriminative, leading to almost perfect performance (e.g. ID 38, 54), but there are cases for which the classification performance is comparable to random guessing.

We observe the distribution of the occurrence of ancestors across different models. Given a patient, for each feature we count the number of times this

Subject ID	unbalance	$\langle BA_{sparse} \rangle$	$\langle BA_{large} \rangle$
1	0.14	0.79(11)	0.77(14)
2	0.31	0.80(6)	0.80(6)
4	0.40	0.80(5)	0.80(6)
5	0.17	0.73(11)	0.73(11)
8	0.26	0.70(13)	0.75(12)
12	0.24	0.87(6)	0.89(7)
16	0.32	0.73(8)	0.73(5)
22	0.12	0.79(16)	0.77(12)
23	0.30	0.81(7)	0.83(8)
24	0.77	0.78(10)	0.80(9)
27	0.12	0.84(16)	0.80(19)
29	0.10	0.58(10)	0.58(10)
36	0.53	0.69(4)	0.73(5)
38	0.60	0.91(4)	0.91(5)
39	0.22	0.81(9)	0.82(12)
16	0.04	0.49(2)	0.59(20)
42	0.49	0.79(10)	0.79(8)
45	0.38	0.78(10)	0.77(9)
49	0.23	0.72(9)	0.74(9)
50	0.40	0.76(4)	0.75(7)
51	0.17	0.79(10)	0.78(9)
53	0.17	0.58(9)	0.63(10)
54	0.44	0.94(5)	0.95(5)
56	0.11	0.70(16)	0.65(10)
57	0.09	0.73(10)	0.75(14)

Table 33: Experiment IV. Balanced accuracy scores from single patient models. We used a nested approach for feature selection on patients from Engel I class. The unbalance refers to the proportion between epileptic contacts over the total. We observe that, for some patients, a linear model with standard clinical descriptors is able to discriminate pathological and physiological recordings. In other cases, the prediction capacity is poor. The difference between the most sparse and the largest model is not significative in terms of performance, as expected in the asymptotical regime.

was selected as an ancestor, across the ten repetitions of the experiment. At the end of this procedure, given a feature, we report the histogram of these occurrences across the population of the Engel I patients, using the box plots as reported in Figure 34. Subgroups (i), (ii), (iii), (iv), and (v) refer respectively to the *mom*, *fft*, *dwt*, *med*, and *thr* features. We observe that, coherently to the

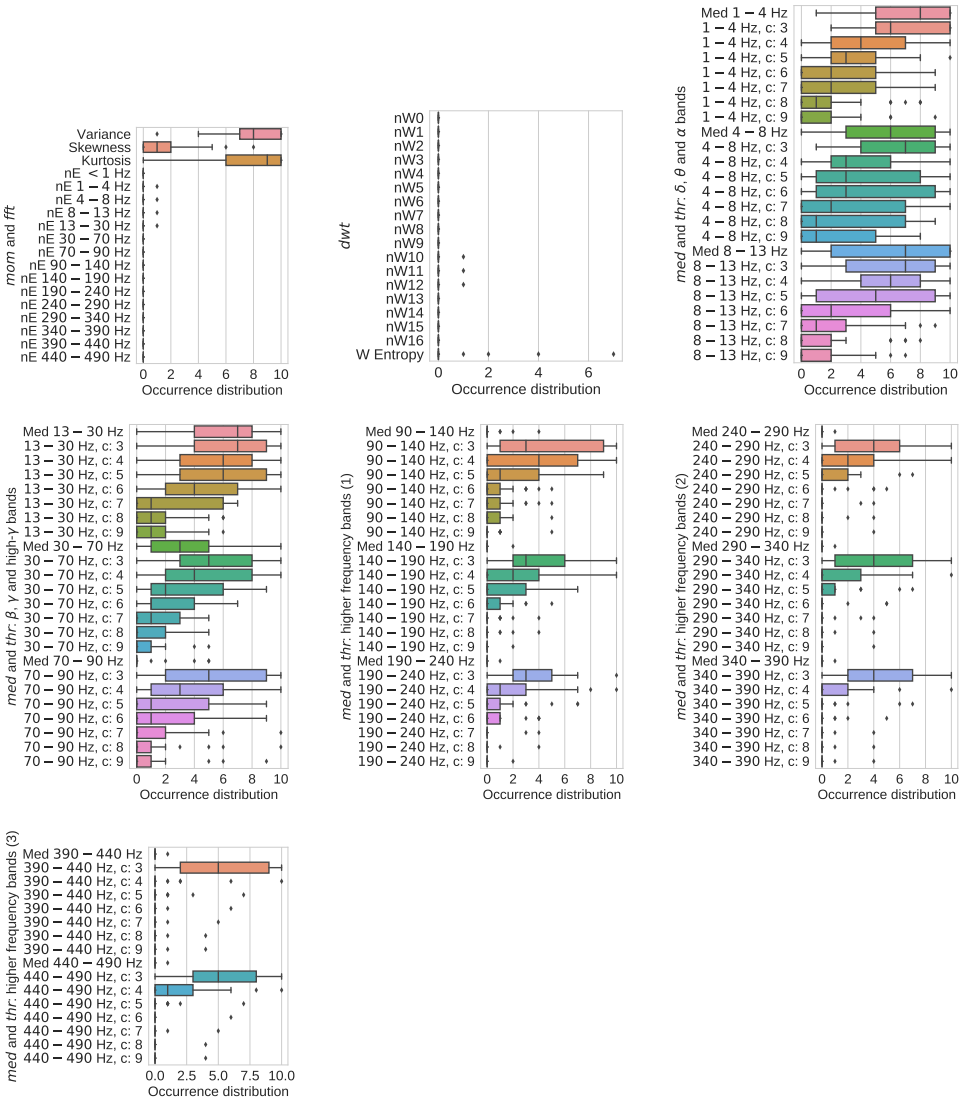

Table 34: Box plots of ancestors occurrence across single patient models for Experiment IV. For each patient we count the number of times each feature has been selected as ancestor across the 10 repetitions of the experiment. We report the distribution of this value across the 25 subjects.

results from Experiment II, features which quantify the mean energy across the ten minutes of acquisitions as *fft* and *dwt* are not relevant to the predictive task. Among the most occurrent ancestors there are variance and skewness of the time series (from the *mom* group). We observe the selection of *med* in concomitance to *thr* features for slow rhythms, δ, α, and β rhythms. As the frequency increases (higher rhythms, $f >$ high-γ), the *med features* are not as meaningful as *thr* features in the same band. The *thr* features are selected as ancestors across repetitions, for all the patients, confirming the generality of results in Experiment III.

7.6 Variability of the Time Series

As last consideration, we attempt at visualizing and quantifying variations in the electrophysiological signals. This, together with the insights from this Chapter, will suggest us future steps. Given that abnormal high amplitude patterns emerged as the most relevant aspects in the discriminative task, we measure here the average variations of the time series from its mean for different time scale. We will observe qualitatively the high variability across subjects.

We consider the outcome of the preprocessing step as described in Section 7.1.1.

Given a patient p, we evaluate the variance of the signal $S_i^{(p)}$ for non overlapping windows in the temporal domain. We repeat this procedure for different window widths w, assuming values in $[1, 10, 30, 50]$ seconds. This operation is equivalent to Formula 94

$$V_w\left(S_i^{(p)}\right)[j] = \int_{j(w\cdot f_s)}^{(j+1)w\cdot f_s} \left(S_i^{(p)}(t) - \mathbb{E}[S_i^{(p)}]\right)^2 dt, \tag{94}$$

where the vector length depends on w. We repeat this procedure for all the contacts and we divide in two subsets the resulting vectors, depending on the presurgical assessment $y_i^{(p)}$, for the $V_w\left(S_i^{(p)}\right)[j]$. We averaged these arrays for the two classes, so to visualize the average variability and its standard deviation at different time scales for contacts in epileptogenic and non epileptogenic zones.

In Figure 43 and Figure 44 we report this result across all the w scales, from the smallest, $w = 1$ s, on top, to the largest, $w = 50$ s, at the bottom. The first pair of figures shows two patients from the Engel I class, the last is relative to Engel IV patients. For all plots the areas and markers denote respectively the standard deviation and mean value of the variability of contacts from the same class, in orange for the positive and green for the negative class.

We observe that, depending on the patient, the discrimination of the positive and the negative samples based on recordings variability emerges as sufficient for some cases, but it is not discriminative in others. To highlight this aspect we report respectively two clinical cases from Engel I class, in Figure 43, and two patients belonging to class Engel IV, in Figure 44.

Starting from class Engel I, for subject 1, in 43, the overlap of the activity generated by the two classes is almost perfect. An evaluation from the time series variability would not be effective in this case. For what regards class Engel IV, in 44, the pre-surgical assessment in subjects 58 and 46 shows that the variability between the two pre-surgical assessed classes is extremely different, despite the bad post-surgical outcome. One among the hypothesis discussed at the beginning of the Section applies for patient 58: the number of thermocoaugulated contacts is much lower that the number of contacts tagged as in the epileptogenic zone from the pre-surgical assessment.

We observe that the variability across different scales w changes depending on the patient, but shorter transients seem to be more meaningful for the discriminative task.

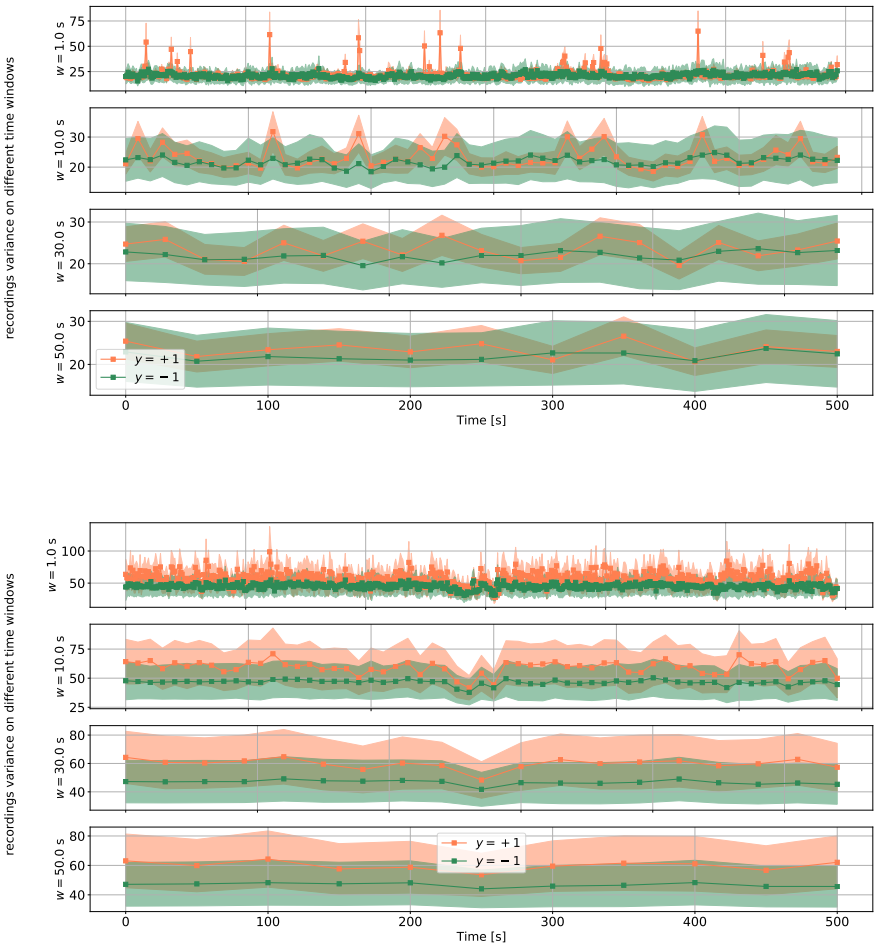

Figure 43: Variability of the time series for patients of class Engel I. Top: subject ID 1, ablated subject with mesial frontal focal epilepsy, #C= 91, #PC= 13. Bottom: subject ID 49, thermo-coaugulated subject, #C= 148, #PC= 34.

Moreover these plots confirm that the over-threshold activity, and consequently the time spent in this regime, heavily depend on the patient.

To get a further insight, we report in Table 35 the values related to the variability of the time series for the two classes. Given a time scale, for each bipolar contact we compute the mean variability across the entire recording and then we average these values across patients, by keeping separated the two classes. The result is a consistent overlap for the two classes. This result suggests that the inference of classification models which take into account the activity from a population does not represent an optimal solution for the classification task.

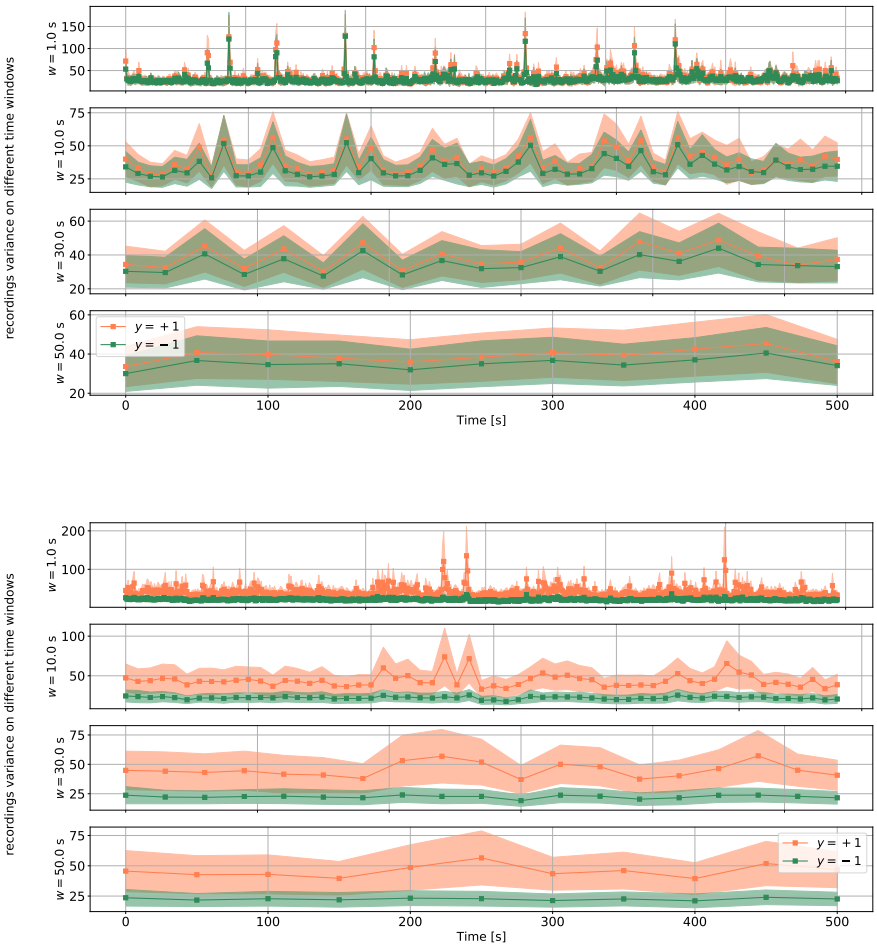

Figure 44: Variability of the time series for patients of class Engel IV. Top: subject ID: 58 , thermo-coaugulated subject with nodular heterotopia, #C= 127, #PC= 43. Bottom: subject ID: 46, thermo-coaugulated subject with right temporal opercular focal epilesy, #C= 143, #PC= 57.

w [s]	$\frac{1}{N}\sum_{p=1}^{n}\left\langle V_w^{(p)}\right\rangle_{y=+1}$ [μV^2]	$\frac{1}{n}\sum_{p=1}^{N}\left\langle V_w^{(p)}\right\rangle_{y=-1}$ [μV^2]
1	53(50)	29(26)
10	57(53)	31(29)
30	58(54)	31(29)
50	59(54)	31(29)

Table 35: The variability of the time series is averaged across the Engel I patients. We observe that mixing all patients is probably not the optimal solution in term of separation between the two classes.

7.7 Comments

Let us summarize the considerations arising from this Chapter in the following, before moving on.

First of all, through a feature selection approach we observed across all the experiments that the collaborative contribution of several bands in the spectrum emerges as key to discriminate the epileptogenic zones. This result does not allow to filter the signal at a specific band while discarding others, so to ease the burden which clinicians must face in the EZ localization task.

Secondly, this result suggests that neural rhythms, as defined in standard neurophysiology, are not optimal in terms of signal interpretability, for the EZ localization task. In this regard, a more appropriate way to proceed would be taking into account more complex waveforms in the neural signals. Several papers move in this direction, focusing of the search of short bursts rather than on the average behavior of the neural signal. As first pointed out by Jasper [64] the typical neural rhythms may be not as significative as specific patterns of the brain. Moreover, variability in the wave patterns may indicate the cooperative activation of particular families of neuronal ensembles and well as neurotransmitters. The work of Voytek and Cole [26] highlights an important aspect related to the hypothesis of Jasper. Indeed the authors observed that several neural measures may be biased by the search of rhythms, while short patterns with a very specific shape (the μ-rhythm is one example) may be the major contributors to the neural activity. In this regard, the recent work of La Tour et al. [71] focuses on signal patterns, with the claim of getting more interpretable results.

Moreover, in the last part of the Chapter, we have been able to give an insight of clinical interest, by performing the experiments on post-surgical seizure free patients. Overall these results do not highlight significative differences in terms of predictive performance when compared to the previous ones. We made the hypothesis that this could result from a not straightforward relationship between the pre-surgical assessment of the epileptic areas and the surgically ablated area.

In Experiment IV we designed models on single patients. Here the classification performance strongly depends on the subject under analysis. The strong contribution deriving from features related to the presence of pattern of abnormal amplitude is consistent with the previous results. This last result shows that for some patients a linear predictive model is sufficient to capture the relation to discriminate the activity recorded in the EZ, but for others it gives balanced accuracy scores comparable to chance level. This result highlights the limited predictive power of the interictal stage when the analysis is limited to the signal amplitude.

Finally, the last qualitative results aimed at measuring the variance of the electrophysiological recordings confirm the importance of building instruments capable of considering single patients, as the discrimination based on abnormal activity may not always be sufficient at the interictal stage.

8

Search of Bursts of Epileptic Activity

In the light of the previous analysis on Engel I patients, we obtained an insight about the crucial role of short patterns of high amplitude for the definition of the EZ. The goal of this Chapter is to illustrate a tool for their search using an open software library, which we adapt to SEEG data. We present some preliminary results which mostly involve data visualization and unsupervised learning techniques for dimensionality reduction and clustering. Given the difficulty in characterizing high amplitude patterns as typical epileptic patterns, we will resort once again on the clinical pre-surgical assessment for patients belonging to the Engel I class. This Chapter opens a plethora of questions and possible future directions to explore, about signal generation and its propagation through different brain regions.

8.1 Short Patterns of Pathological Activity

From previous analysis we obtained conclusive results of classification methods trained on interpretable features, extracted using signal processing techniques. In this regard, features which take into account the neural activity above baseline, with strong contributions from several frequency bands in the spectrum, emerged as relevant. Electrophysiological patterns potentially related to the EZ may manifest with high variability of their temporal profile due to (i) the area involved in their generation; (ii) distortions due to anisotropy of the neural tissue, as nearby neural populations may activate; (iii) different pharmacological treatments which may impact on inhibitory/ excitatory mechanisms related to signal propagation. In this view, the identification of epileptic patterns and the following characterization may get extremely challenging. Moreover, the capacity of finding epileptic patterns using data driven tools may be strongly affected by the rarity of these events across the neural recordings during the interictal stage.

In the initial Chapters we debated extensively about candidate patterns of epileptic activity at the interictal stage. Given the presence of all the frequency components, we decide to focus here on patterns which we hypothesized to be interictal epileptic spikes. Our candidate epileptic patterns will be extracted at relatively low frequency, using a wide band-pass filter. We do not exclude

the possibility of applying the methodology proposed below to any other frequency range of interest (e.g. HFOs).

In this regard, some machine learning tools developed for interictal spike detection [93, 104] mostly deals with large scale electrophysiology and are applicable on signal acquired from scalp. Given these patterns, it is nonetheless controversial their definition as truly epileptic spikes. Several studies highlight the disagreement among medical experts about the epileptic or physiological nature of such patterns [9, 66].

On the other hand, algorithms for detection and analysis of spike from single as multiple neurons exist [91, 134]. The spatial resolution hypothesized here is much higher given the different type of measure. In the following we leverage on one among those, Spyking-Circus [134], an open-source Python library whose goal is spike sorting for signal recorded using multi-electrode arrays. The method has been designed for the search of spikes from thousands multi-electrodes recordings, with particular attention to the computational performance. As the algorithm is well-cited and the library seems maintained throughout the last years, we rely on it for our analysis. Reviewing the literature related to spike detection, we notice the use of common strategies across these methods, as signal thresholding, PCA, and clustering.

8.2 Spyking-Circus Specifics

The method leverages not only on the temporal nature of the data, but also on the spatial position of each recording. Based on this a priori co-occurrent patterns recorded at different positions may derive from the propagation of a common signal, or may be triggered by the activation of a third common signal generator. The spatial prior allows to infer more reliable waveforms of spike activity. The original domain of application of this tool is micro-electrode recording systems, which acquire local voltage from separated cellular sites, at a distance of micrometers.

Even in the best case scenario, for neighborhood contacts, the SEEG contacts distance are orders of magnitude above the micrometer scale. In our limited understanding, this does not represent a limit to the Spyking-Circus performance for pattern detection. The algorithm may have difficulties in grouping together patterns due to the prior knowledge about high distance among contacts, but this should not impair dramatically the identification and localization of templates in time.

8.2.1 *The Algorithm*

The method is characterized by different phases. We summarize its essential steps and put more emphasis in the description on those steps and hyperparameters which depend on the users choice. We suggest a further reading of the original work to get a clearer view of the method.

Guided by our task we slightly adapt the algorithm in some part to work on SEEG recordings. Notice that for this analysis we will process one patient at the time.

1. INPUT DATA (i) a file containing the arrays on neural recordings, with dimensions $\#\mathcal{C}^{(p)} \times N_T$; (ii) a related file, containing the contacts positions, given in μm; (iii) a third file containing the main settings of the algorithm (see Table 1 in [134]).

2. PREPROCESSING by default it consists in (i) data filtering, (ii) spike detection, (iii) whitening, and (iv) basis estimation. (i) By default the method applies a high-pass Butterworth filter, where low and high cut-off frequencies must be specified. (ii) At this step, for each contact k, the algorithm computes a spike threshold based on Median Absolute Deviation (MAD) of the signal, as $\theta_k = \lambda \text{MAD}(S_k^{(p)}(t))$, with λ free parameter. A candidate spike time is the one which exceeds the θ_k value. Positive, negative or both signs thresholds can be imposed. (iii) During whitening spurious spatial correlations are removed. (iv) Dimensionality reduction through PCA for the set of collected waveforms. After having aligned all the waveforms, the algorithm up-samples the patterns, we reduce to a number of components equivalent to 5.

3. CLUSTERING the goal here is finding a dictionary of waveforms. The main steps consist here in (i) masking, (ii) pre-clustering, (iii) clustering by local density peak, (iv) centroids and cluster definition, (v) clusters merge, (vi) templates definition, and (vii) removal of redundant templates. We do not go in detail here. The algorithm clusters the templates both in time (e.g. patterns recorded from the same channels across different time instants) and space (e.g. co-occurrent patterns recorded from different electrodes). The main parameter here is ρ, which denotes the closest contact distance. The output of the clustering step is a dictionary of non redundant, one dimensional templates, each of shape N_t. The output of this operation is the set of K templates, $\{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$, with a number of elements depending on the data.

4. TEMPLATE MATCHING at this step the algorithm leverages on an iterative greedy approach to estimate the putative spike times. The output of this operation is a set of K lists of different sizes $\{[T_1^1, \ldots T_{t_1}^1], \ldots, [T_1^K, \ldots T_{t_K}^K]\}$, where the generic list k contains the times at which the $\mathbf{x}_k$ pattern has been detected.

8.2.2 *Parameters Choice and Small Modifications*

Here we recall the previous steps and we illustrate our choice for the analysis.

1. For each patient we give as input to the method a matrix of already preprocessed SEEG recordings. These recordings are results of a band-passed with a FIR filter (Hanning window) [48], with low and high cut-off frequencies of 1

Hz and 40 Hz respectively, and low and high transition bandwidths equivalent to 1 and 10 Hz respectively. As at a preliminary stage, we were interested in detecting only the highest amplitudes and from the previous results, a reasonable choice is to set the cut off frequency at 40 Hz, in order to avoid line effects. The second input file must contain the contacts position. Spyking-Circus was designed for 2-dimensional probes only, so we slightly modify it to our geometry. We pass to the algorithm the SEEG positions in the 3D Euclidean coordinates, as given by FreeSurfer. In Table 36 we report the main setting of the configuration file. We fixed ρ to the value of 250 μm, which is a slight

parameters	cut-off	polarity	λ	N_t	$\rho[\mu\text{m}]$
	$[2, 40]$	both	$[7, 9]$	1001	250

Table 36: Our setting for the Spyking-Circus parameters. Filter refers to the preprocessing which the algorithm performs internally, λ to the threshold constant, N_t to the number of time points for each template.

under-estimate of the distance for contacts on the same electrode.

2. Our choice regarding (ii) was to detect patterns with both positive and negative polarities. This is indeed a common scenario in the analysis of SEEG recordings. As an example we report in Figure 45 two templates, extracted by Spyking-Circus from the same patient.

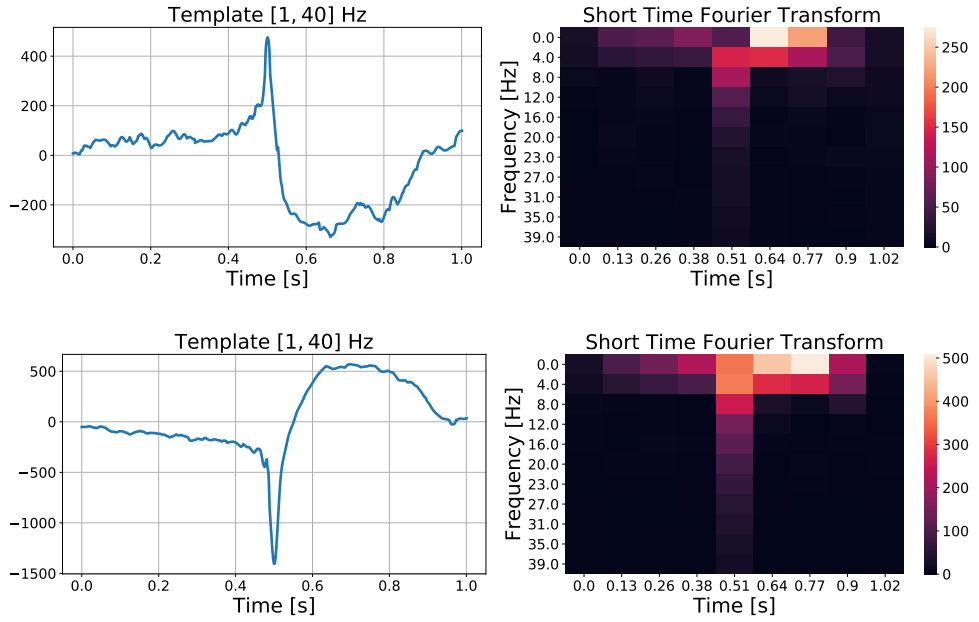

Figure 45: Given a patient, we ran the Spyking-Circus algorithm. We show here two templates, similar in shape but of opposite polarity. The blue curves correspond to the template in the temporal domain. On the right, we report the absolute value of the Short Time Fourier Transform, for the two templates. We observe that the sharp central peaks have effect across all frequencies, while the slow waves reflects on the low frequencies only.

8.3 Search for Common Activity

With the goal of searching for any common signature of the epileptic activity across patients, as a first step we compare templates extracted from different patients. We decide to test this hypothesis as we notice again high variability of the extracted waveform shapes across patients. We report in Figure 46 a template for a different patient from the one who generated the activity shown in Figure 45.

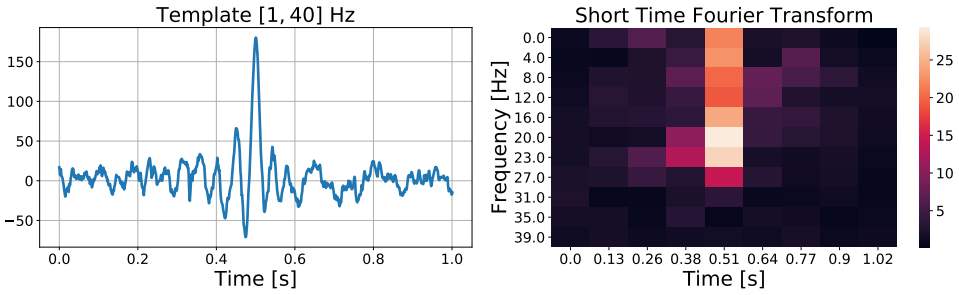

Figure 46: The patterns variability among patients is extremely high, as a qualitative comparison with waveforms in Figure 45 shows.

In addition to the different shape, also strong differences in amplitude are present. To reduce their impact on the result we leverage of a distance which takes into account the profile of the temporal patterns, rather than their amplitudes. We measure then pairwise similarity across patterns using the cosine metrics in Formula 95

$$d_{\text{cosine}}(x, y) = 1 - \frac{|\langle x, y \rangle|}{\|x\| \|y\|}. \tag{95}$$

We introduce in the measure the absolute value to avoid the dependence from the spike polarity, which, given our choice of considering both, may flip sign.

8.3.1 *Comparison Across Patients*

From this analysis we aim at observing if the templates, as extracted from Spyking-Circus, group together in relation to the patient or if they are almost independent from their domain (the subject in this case). If this dependence does not hold true, it would justify a further search of common short time epileptic activity. To evaluate if this is the case we rely on the metric of optimality defined for clustering, in particular the silhouette score.

To proceed in the task, we resort to the absolute cosine distance in Eq. 95 to first measure the similarity among patients. We leverage on an agglomerative clustering approach[1] which takes as input the precomputed matrix of pair-

1 Last access: October 28th, 2019 `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html#sklearn.cluster.AgglomerativeClustering`

wise distances. We set our choice to complete linkage. We vary the number of possible clusters in the range #clusters $\in \{2, \dots, 40\}$. We give as input to the learning algorithm 80% of the templates randomly extracted with replacement from the entire dataset, which consists of a total of 1488 templates for the 25 patients of class Engel I.

We repeat the experiment 10 times. For each repetition, at each #clusters we measure the silhouette score. The result of this procedure is report in Figure 47. On the x-axis we report #clusters, the number of clusters, on the y-axis the

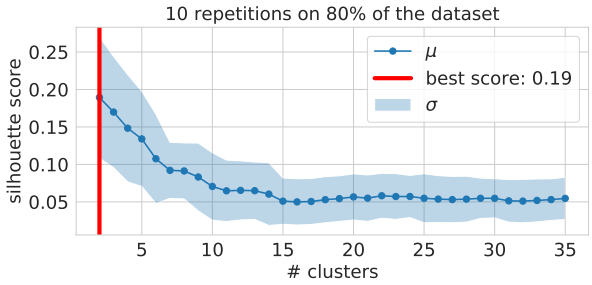

Figure 47: The curve of silhouette score across 10 repetitions of the experiment. For each repetition we give 80% of the dataset as learning set. The blue curve and area represent respectively the mean and the standard deviation of the silhouette score for different #clusters values. The red line denotes the best #clusters value, correspondent to the best mean silhouette score.

silhouette score. The blue curve and the blue area correspond respectively to the silhouette mean and standard deviation across repetitions. The vertical red line denotes the $(\#\text{clusters})^* = 2$ for which we obtain the best mean score.

Given the choice of $(\#\text{clusters})^*$, we refit the entire dataset. In Figure 48 we report the two centroids. This result is promising in terms of comparison

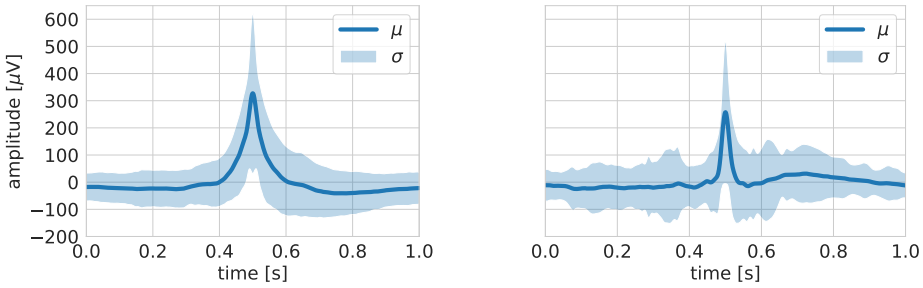

Figure 48: We refit the agglomerative clustering method, imposing the $\#(\text{clusters})^* = 2$. We report the two centroids. x-axis: time axis, in seconds, y-axis: amplitude.

across patients. We notice that the two centroids are characterized by a sharp central peak, of time width 200 ms. We observe nonetheless the large variance of these temporal wave profiles. To more light on the epileptic spikes, we want to observe if large amplitude patterns recorded in epileptogenic areas tends to have a more defined shape, when compared to signals recorded from other non epileptogenic areas. For this task tagged patterns are needed.

8.3.2 *Tagging Patterns*

Here we aim at observing if spikes generated from the EZ tend to aggregate more compactly than sharp patterns which manifest in the non epileptogenic zones (non EZ). For this scope we need to tag temporal patterns as belonging or not to epileptogenic zones. We decide to exclude for all patients those templates which overall occur less than three times across the entire acquisition. This choice may look arbitrary but it should reduce the amount of noise, as we risk to discard extremely rare templates but also random fluctuations and possible signal artifacts.

The tag assignment is performed separately for each patient, as shown in Figure 49.

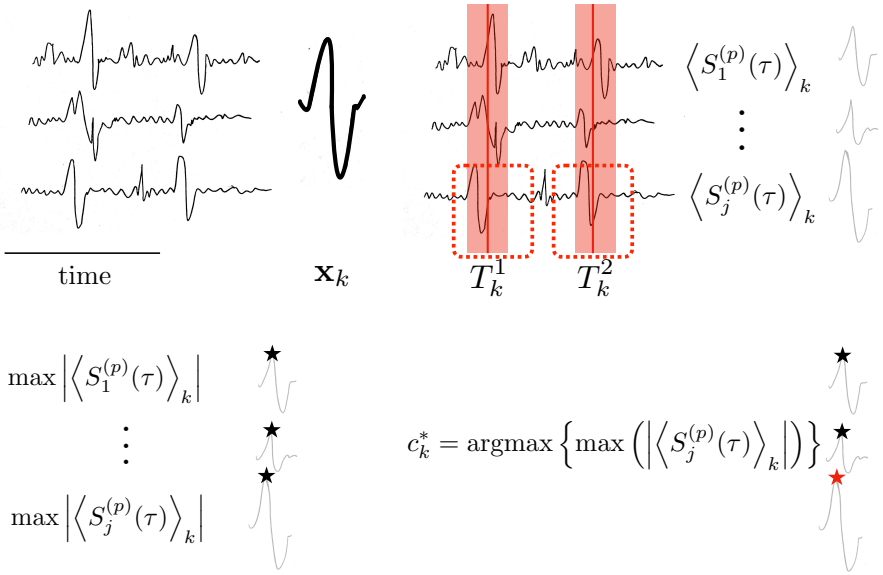

Figure 49: From the top left, using Spyking Circus we extract for each patient a set of templates. We show the workflow for a generic template $\mathbf{x}_k$. The other output of the algorithm is the time instants corresponding to $\mathbf{x}_k$. For each contact we compute the average of these temporal chunks, highlighted in red. This operation gives as result a pattern of 1 s length for each contact. For each of those we compute the maximum amplitude, by taking the absolute value of the signal, as in the bottom left. The label is assigned by considering the channel which produces the higher activity.

Given recordings from the p-th subject, we consider one pattern $\mathbf{x}_k$ at the time. We average the activity on each contact, for those time windows where the template $\mathbf{x}_k$ has been detected. This corresponds to the Formula 96

$$\left\langle S_j^{(p)}(\tau)\right\rangle_k = \frac{1}{t_k}\sum_{s=1}^{t_k} S_j^{(p)}(T_s^k + \tau), \text{ with } \tau \in [-0.5, 0.5] \text{ s.} \tag{96}$$

The output of this operation is the average activity across time windows which match to the template $\mathbf{x}_k$, a pattern of duration 1 second, for each recording.

In order to assign a label to the template k for the p-th patient, we first evaluate at each contact the maximum amplitude, in its absolute value. This

operation corresponds to the bottom left part in Figure 49. Finally we consider which channel produces the maximum overall activity. We call this contact c_k^* and we assigned to the template $\mathbf{x}_k$ its label, as in Eq. 97.

$$c_k^* = \arg\max_c \left\{ \max \left(\left| \left\langle S_j^{(p)}(\tau) \right\rangle_k \right| \right) \right\} \tag{97}$$

$$y^{(p)}_{\text{template } k} = y^{(p)}_{c_k^*}. \tag{98}$$

The hypothesis is here of instantaneous propagation, so that the maximum amplitude implies minimum distance to the area generating the activity. Our labeling procedure gives origin to two subsets of templates

#events with maximum amplitude in EZ = 461,

#events with maximum amplitude in non EZ = 566.

8.3.2.1 *Analysis*

To evaluate if interictal spikes have a more similar waveform among them than other high amplitude events we resort to a recent approach aimed at detecting HFOs [78]. Here the authors, after collecting HFO events, show that the epileptic patterns tends to be more similar in shape than other high frequency waveforms above threshold. We implement a similar strategy for the interictal spikes. The analysis leverages on DBSCAN [39], an unsupervised method which performs data clustering. We resort to the scikit-learn implementation of DBSCAN[2].

For this experiment we use the tag assignment as in Equation 97. We run separately the DBSCAN algorithm on the two classes: patterns in the EZ against patterns in the non EZ. Given the different amount of templates with positive and negative tags, we normalize the number of clusters by dividing the output of the method for the number of elements from a class. We give as input to DBSCAN a normalized version of the templates, with positive polarization. We use the standard metric of the experiment in [78], which is the Euclidean norm, supposing that this is the best metric to use. In this concern, since we imposed the unit norm for the templates, the Euclidean distance corresponds to the square root of the cosine distance. We do not expect high difference between the two. We let the tolerance ray of DBSCAN vary in the range $[10^{-3}, 1]$. If the hypothesis of similarity for interictal spikes is verified, we expect the number of clusters for these patterns to collapse to one already for small ray values.

We report in Figure 50 the curve related to the normalized number of clusters as function of the ray parameter ε of DBSCAN. The results are not encouraging, as the *EZ* curve is comparable to *non EZ* curve. In the next section we propose a further attempt to improve this result, which consists in the imposition of prior knowledge on the spike waveform.

2 Last access: October 28th, 2019 `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html`

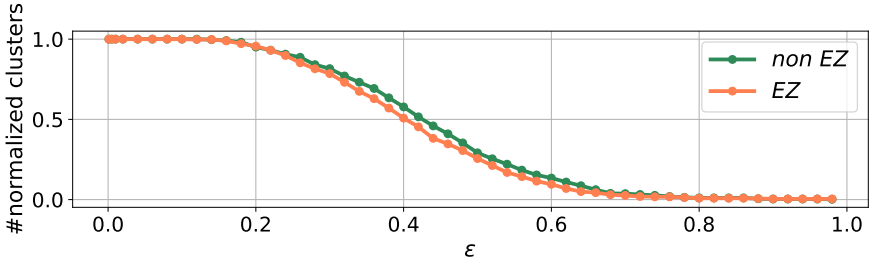

Figure 50: Curve of normalized number of clusters as function of ε. We perform the clustering separately for the two classes. In green: *non EZ* clustering results for events with maximum amplitude in non EZ; in orange: *EZ* clustering results for events with maximum amplitude in EZ. In case of high similarity among interictal spikes we would have observed a drop in the curve. The two curves are almost indistinguishable.

8.3.2.2 *Prior Imposition on the Waveform*

To select the patterns we require each template to satisfy some a priori. The criteria used to select the candidate spikes templates are related to the work of Latka et al. [74].

Given a template we used different representations of the patterns through wavelet transform as a way to impose some selection criteria. We chose both discrete and continuous mother wavelets, in particular

(i) Mexican hat wavelet;

(ii) complex Morlet wavelet;

(iii) Daubechies of order 4^{th}.

For what regards (i) and (ii), the scales are set so to share the same central frequency values. We then discard frequencies above 40 Hz. Our knowledge imposition on the patterns shape is extremely mild. Given that for each template the maximum amplitude is centered in the temporal interval, we impose *concentration*, *asymmetry*, and *decreasing amplitude*.

CONCENTRATION The energy of the wavelet coefficients must be mostly concentrated in the center of the window of one second length. We consider the Daubechies wavelet and we divide the temporal dimension in three parts of equal size. We compute the absolute value of the detail coefficient. Given a scale we sum up the coefficients from the three windows of equal size $(T^1_{1/3}, T^2_{1/3}, T^3_{1/3})$. We require the value for the central window to be higher than the other two. We perform this operation for the first three scales.

$$\sum_t cD[t \in T^1_{1/3}] < \sum_t cD[t \in T^2_{1/3}] \text{ and } \sum_t cD[t \in T_3] < \sum_t T^3_{1/3} cD[t \in T^2_{1/3}]$$

We perform a similar operation on the representation from the Morlet mother wavelet. Given that the peak usually has a width of 200 ms, at a fixed scale we sum the absolute values of all the coefficients in the first 400 ms, between 400

and 600 ms, and in the last 400 ms. We call the three windows $\mathcal{T}_1, \mathcal{T}_2$, and $\mathcal{T}_3$. We normalize the sums based on the temporal interval lengths. We require for the central value to be greater than the left and right sides values.

$$\frac{1}{\#\mathcal{T}_1}\sum_{\tau\in\mathcal{T}_1}\sum_{a}|C(\tau,a)| < \frac{1}{\#\mathcal{T}_2}\sum_{\tau\in\#\mathcal{T}_2}\sum_{a}|C(\tau,a)|,$$
$$\frac{1}{\#\mathcal{T}_3}\sum_{\tau\in\mathcal{T}_3}\sum_{a}|C(\tau,a)| < \frac{1}{\#\mathcal{T}_2}\sum_{\tau\in\#\mathcal{T}_2}\sum_{a}|C(\tau,a)|, \text{ with } C \text{ Morlet coefficients.}$$

ASYMMETRY The asymmetry of the patterns must reflect on the coefficients distribution, in case of symmetric mother wavelets. Epileptic spikes are indeed usually followed by slow waves. We verify this requirement using the Mexican hat transform. We consider the absolute value of the coefficients across scales. This operation returns an array of length $N_T = 1001$ points. We sum the values for $T^1_{1/2} = t < 0.5$ s and $T^2_{1/2} = t \geq 0.5$ s. We require this last value to be greater than the former,

$$\sum_{\tau\in T^1_{1/2}}\sum_{a}|C(\tau,a)| < \sum_{\tau\in T^2_{1/2}}\sum_{a}|C(\tau,a)| \text{ with } C \text{ Mexican hat coefficients.}$$

DECREASING AMPLITUDE The signal should have coefficients different from zero at all scales, with a decrease in the energy as the frequency increases. We use coefficients from the complex Morlet wavelet transform. We divide the central frequencies in different bands: $b_1 = 1-5$Hz, $b_2 = 5-8$ Hz, $b_3 = 8-12$ Hz, $b_4 = 12-15$ Hz, $b_5 = 15-20$ Hz, $b_6 = 20-30$ Hz, and $b_7 = 30-40$ Hz. We sum the absolute value of the coefficients across time for each scale. We then sum the outcome vector over all the scales included in each interval. For this vector we require to measure smaller entries as the frequency band values increase

$$\sum_{a \text{ s.t. } f^a_c\in b_k}\sum_{i=1}^{T}|C(\tau,a)| \geq \sum_{a \text{ s.t. } f^a_c\in b_{k+1}}\sum_{i=1}^{T}|C(\tau,a)|, \qquad \forall k\in\{1,\dots,6\},$$

with C Morlet coefficients.

In Figure 51 we report three different examples of the wavelet criteria. Starting from the left, we impose the asymmetry of the wavelet coefficients with respect to the center of the time interval. In the middle, we show the prior related to energy concentration for the complex Morlet wavelet. The same a priori is verified by the pattern on the right, using the Daubechies wavelet.

We report in Table 37 the results of the selection.

As we can notice, the ratio between the number of event tagged as in the EZ and the ones in the non EZ increases based on wavelet criteria, but the approach is still insufficient to group efficiently the epileptic spikes. In the same fashion on the previous analysis, we report the result of the DBSCAN fit on the selected events, in Figure 52. On top we show the previous result before selection, while at the bottom, the plot reports the curves from DBSCAN algorithm after selecting events using the wavelet criteria. This approach does not lead to effective results for the identification of interictal spikes.

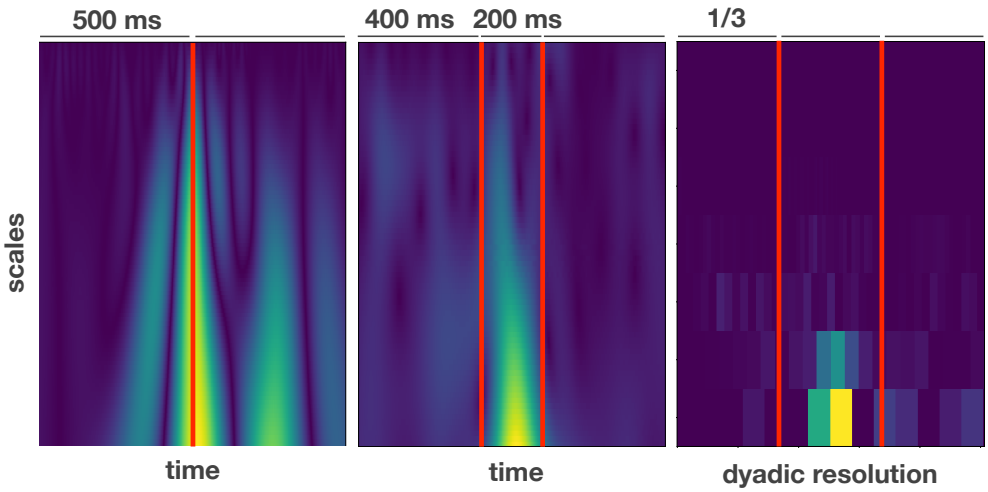

Figure 51: Absolute value of wavelet coefficients computed for a candidate epileptic pattern. Starting from the left to the right: Mexican hat wavelet, complex Morlet, and Daubechies 4^{th}. The first two plots report on the x-axis the time domain, each of length 1 second. On the y-axis wavelet scales. Starting from the left we impose asymmetry of the coefficients distribution, in the middle and right plots we impose coefficients concentration.

	#events EZ	#events non EZ	(#events EZ)/#events
before selection	461	566	0.49
after selection	266	207	0.56

Table 37: Results of selection of templates based on prior knowledge. The entries #events EZ and #events non EZ denote respectively the number of events for which the maximum amplitude of the average is recorded from a epileptogenic zone or to a non epileptogenic zone. We observe a reduction in the number of non epileptogenic patterns.

8.4 Comments

Given the results at hand it is hard to proceed in further analysis for the identification of common patterns of activity in the interictal period in the range $[1, 40]$ Hz by this approach. The interpretation of these results is not straightforward, given the absence of a ground-truth about the tag assignment. The assumption of a single pathological pattern as an approximation to the common pathological activity to the EZ in the range $[1, 40]$ Hz could be too strong. Moreover the attempt of aggregating pathological templates independently from the region of origin may represent an over optimistic approach. About the last point, the optimal labeling would arise from considering the region which generates the pattern, rather than the maximum amplitude approach in

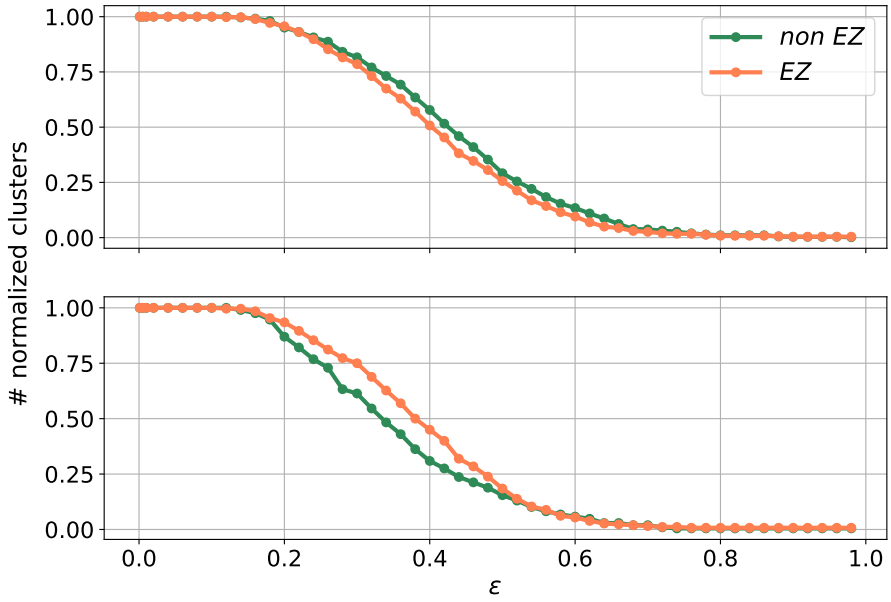

Figure 52: The curve of the normalized number of clusters as function of the ray parameter in DBSCAN. The orange curve shows the results for patterns tagged as from the EZ, while in green we report the ones in the non EZ. On the top, we report the curve before applying the wavelet selection criteria, on the bottom the curves are computed for the templates that pass the selection criteria.

Eq. 97. At this point these considerations represent conjectures which require a future validation.

PART IV

Conclusions

9

Considerations and Future Works

9.1 Summary

In this work we tackled the analysis and characterization of focal epilepsy, a pathological condition which manifests through convulsions, states of impaired consciousness, impaired limbs mobility, auras, déjà-vu, causing a degradation of the quality of life.

We started by considering the clinical state-of-the-art, and focusing mostly on electrophysiological signals acquired from the brain. The goal of automatic characterization of the signal in the EZ, aimed at its localization, is extremely challenging. This is mainly due to the abundance of candidate biomarkers encountered in this context, the clinical validation needed for this type of study, and the general difficulty of reaching perfect agreement among experts.

A prototypical example of this situation is the HFO pattern, which we reviewed in greater detail. Different articles in this context present apparently contradictory results, with some authors claiming the importance of HFOs for the EZ localization, and others discarding this hypothesis.

In the light of several metrics used across these papers, we suggest that HFOs may be insufficient to localize the entire EZ, due to the high number of EZ regions which do not manifest HFOs. Nonetheless, HFOs may still potentially be a good tool for the identification of the EZ, given the low number of false positive regions (presence of HFOs in the non EZ area) especially when the presence of HFOs in assessed by multiple repetitions of these events during the acquisition.

In this thesis we instead consider all the candidate EZ biomarkers in the electrophysiological signal, rather than focusing on a specific frequency, by basing our analysis on machine learning and regularized techniques. Through these tools we aim at

(i) revealing the most salient aspects of interest for the discrimination of the EZ from the physiological area without the imposition of prior knowledge;

(ii) classifying SEEG recordings, as acquired from the EZ or the non EZ areas.

For this scope, we performed several analyses on a population consisting of 60 patients, presented extensively in Chapter 4. The analysis was mostly focused on the electrophysiological signal acquired through invasive pre-surgical SEEG acquisition system.

Chapter 5 represents a preliminary analysis of the feasibility of this study, giving a positive perspective about the predictive role of automatic approaches for the localization of the EZ, based on features commonly defined in the clinical literature. This analysis takes advantage of data representation, signal processing, and feature extraction to classify chunks of interictal activity. We underline that this approach comes with several limitations, due to the learning protocol which we tackle later on in Chapter 7.

Chapter 6 presents MT-MKL, a machine learning tool for the analysis of neural recordings. This allowed us to characterize the activity of the entire population of epileptic patients, by leveraging to a multiscale decomposition of the entire time series during the interictal stage. Following signal decomposition, the method integrates signals at different scales by selecting the most relevant ones for the classification task, based on a regularized multiple kernel learning approach. By taking advantage of sparsity constraints, MT-MKL should provide a measure of the importance of frequency bands common to the entire population, and possibly give an insight about the activity generated in the epileptogenic zones. In a series of experiments we have however highlighted some of its limitations. Retrospectively, the method depends on the similarity of the neural activity among brain regions to assess the EZ, but this does not allow us to distinguish the pathological activity from the physiological one. In particular we observed that further analysis does not lead to any selection of a sparse subset of frequency bands. Nonetheless, from a machine learning perspective the method can still support clinicians in the identification of EZ areas based on similar electrophysiological activity among recordings.

In Chapter 7 we address some limitations from the previous analysis. We proposed a strict learning protocol which does not allow any type of contaminations among data splits. The experiments rely on multiple hand-crafted features which make an extensive use of signal processing techniques and clinical knowledge. Through a series of machine learning experiments, we identified the most promising pathological signatures of the epileptogenic zones. For this last part analyses we got access to the post-surgical assessment, which reduced the dimension of our population but allowed us to get results of clinical relevance.

The main result from the last part of Chapter 7 suggests to discard the automatic analysis of the time series in favor of short patterns of abnormal amplitudes, at several frequency bands. In this regard, Chapter 8 represents a preliminary attempt at identifying relevant patterns of short duration across the neurophysiological recordings, by considering one patient at a time. At this preliminary stage, the approach could provide clinical support for the visualization of candidate pathological activity.

9.2 Future Directions

Among many possible questions and future directions, we delineate a feasible path arising from the last results on the Engel I patients. In Chapter 7 we presented indeed the result of the learning pipeline for feature selection using standard evaluation metrics in machine learning. However, for the task at hand these metrics may lack of interpretability. These scores indeed take into account the number of misclassified samples but we argue that the misclassification may be more or less severe by the distance to the EZ areas. For example, a recording labeled as false positive should be weighted differently depending on whether it was acquired in proximity of the EZ or in a region far away from it. Quantifying the error through this measure may potentially alleviate the issue of evaluating the goodness of such models, which has been a concern given the signal propagation. An analysis left for future research is to use the models trained on Engel I class to classify the activity of Engel IV patients. Here we cannot validate the goodness of the prediction, but we do hypothesize a decrease in performance.

Designing tools for the identification of epileptic patterns seems a promising, despite if difficult, approach. Indeed, the main limitation here is the lack of a ground-truth about the real epileptogenicity of the patterns. Again, we may rely on clinical support, but this would not represent a solution, for the moon-shot goal of automation. We suggest that the missing component critical for the clinical interpretation of results is the lack of prior knowledge about brain structure in our machine learning algorithm.

In general we strongly argue that further investigations of the interictal activity is useful for the characterization of the pathology. This holds particularly true if we wish to automatize the analysis and give support to medical experts throughout the entire data acquisition period, which spans the range of days rather than minutes of signals, and whose complete characterization becomes clearly unfeasible without automatic tools.

Bibliography

[1] Hervé Abdi and Lynne J Williams. 'Principal component analysis.' In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

[2] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. 'Discrete cosine transform.' In: *IEEE transactions on Computers* 100.1 (1974), pp. 90–93.

[3] Naomi S Altman. 'An introduction to kernel and nearest-neighbor nonparametric regression.' In: *The American Statistician* 46.3 (1992), pp. 175–185.

[4] Michael Aminoff, David Greenberg, and Roger Simon. *Clinical neurology*. McGraw-Hill Education, 2015.

[5] Senjian An, Wanquan Liu, and Svetha Venkatesh. 'Face recognition using kernel ridge regression.' In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2007, pp. 1–7.

[6] Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. 'On invariance and selectivity in representation learning.' In: *Information and Inference: A Journal of the IMA* 5.2 (2016), pp. 134–158.

[7] Nachman Aronszajn. 'Theory of reproducing kernels.' In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.

[8] GF Ayala, M Dichter, RJ Gumnit, H Matsumoto, and WA Spencer. 'Genesis of epileptic interictal spikes. New knowledge of cortical feedback systems suggests a neurophysiological explanation of brief paroxysms.' In: *Brain research* 52 (1973), pp. 1–17.

[9] Daniel T Barkmeier, Aashit K Shah, Danny Flanagan, Marie D Atkinson, Rajeev Agarwal, Darren R Fuerst, Kourosh Jafari-Khouzani, and Jeffrey A Loeb. 'High inter-reviewer variability of spike detection on intracranial EEG addressed by an automated multi-channel algorithm.' In: *Clinical Neurophysiology* 123.6 (2012), pp. 1088–1095.

[10] Quentin Barthélemy, Cedric Gouy-Pailler, Yoann Isaac, Antoine Souloumiac, Anthony Larue, and Jérôme I Mars. 'Multivariate temporal dictionary learning for EEG.' In: *Journal of neuroscience methods* 215.1 (2013), pp. 19–28.

[11] Fabrice Bartolomei, Patrick Chauvel, and Fabrice Wendling. 'Epileptogenicity of brain structures in human temporal lobe epilepsy: a quantified study from intracerebral EEG.' In: *Brain* 131.7 (2008), pp. 1818–1830.

[12] Yoshua Bengio. 'Deep learning of representations for unsupervised and transfer learning.' In: *Proceedings of ICML workshop on unsupervised and transfer learning*. 2012, pp. 17–36.

[13] James Bergstra and Yoshua Bengio. 'Random search for hyper-parameter optimization.' In: *Journal of machine learning research* 13.Feb (2012), pp. 281–305.

[14] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. 'Proximal alternating linearized minimization for nonconvex and nonsmooth problems.' In: *Mathematical Programming* 146.1 (2014), pp. 459–494. ISSN: 1436-4646. DOI: 10.1007/s10107-013-0701-9. URL: https://doi.org/10.1007/s10107-013-0701-9.

[15] Poomipat Boonyakitanont, Apiwat Lek-uthai, Krisnachai Chomtho, and Jitkomut Songsiri. 'A review of feature extraction and performance evaluation in epileptic seizure detection using EEG.' In: *Biomedical Signal Processing and Control* 57 (2020), p. 101702.

[16] Karsten M Borgwardt. 'Kernel methods in bioinformatics.' In: *Handbook of statistical bioinformatics*. Springer, 2011, pp. 317–334.

[17] Pierre Bourdillon, Jean Isnard, Hélène Catenoix, Alexandra Montavont, Sylvain Rheims, Philippe Ryvlin, Karine Ostrowsky-Coste, François Mauguiere, and Marc Guénot. 'Stereo electroencephalography–guided radiofrequency thermocoagulation (SEEG-guided RF-TC) in drug-resistant focal epilepsy: Results from a 10-year experience.' In: *Epilepsia* 58.1 (2017), pp. 85–93.

[18] Anatol Bragin, Jerome Engel Jr, Charles L Wilson, Itzhak Fried, and Gyorgy Buzsáki. 'High-frequency oscillations in human brain.' In: *Hippocampus* 9.2 (1999), pp. 137–142.

[19] Leo Breiman. 'Random forests.' In: *Machine learning* 45.1 (2001), pp. 5–32.

[20] Andreas Bruns. 'Fourier-, Hilbert-and wavelet-based signal analysis: are they really different approaches?' In: *Journal of neuroscience methods* 137.2 (2004), pp. 321–332.

[21] Sergey Burnos, Birgit Frauscher, Rina Zelmann, Claire Haegelen, Johannes Sarnthein, and Jean Gotman. 'The morphology of high frequency oscillations (HFO) does not improve delineating the epileptogenic zone.' In: *Clinical Neurophysiology* 127.4 (2016), pp. 2140–2148.

[22] Francesco Cardinale, Massimo Cossu, Laura Castana, Giuseppe Casaceli, Marco Paolo Schiariti, Anna Miserocchi, Dalila Fuschillo, Alessio Moscato, Chiara Caborni, Gabriele Arnulfo, et al. 'Stereoelectroencephalography: surgical methodology, safety, and stereotactic application accuracy in 500 procedures.' In: *Neurosurgery* 72.3 (2012), pp. 353–366.

[23] S Grace Chang, Bin Yu, and Martin Vetterli. 'Adaptive wavelet thresholding for image denoising and compression.' In: *IEEE transactions on image processing* 9.9 (2000), pp. 1532–1546.

[24] Francois Chollet. *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*. MITP-Verlags GmbH & Co. KG, 2018.

[25] Ivan Cohen, Vincent Navarro, Stéphane Clemenceau, Michel Baulac, and Richard Miles. 'On the origin of interictal activity in human temporal lobe epilepsy in vitro.' In: *Science* 298.5597 (2002), pp. 1418–1421.

[26] Scott R Cole and Bradley Voytek. 'Brain oscillations and the importance of waveform shape.' In: *Trends in cognitive sciences* 21.2 (2017), pp. 137–149.

[27] Patrick L Combettes and Băng C Vũ. 'Variable metric forward–backward splitting with applications to monotone inclusions in duality.' In: *Optimization* 63.9 (2014), pp. 1289–1318.

[28] Thomas H Cormen. *Introduction to algorithms*. MIT press, 2009.

[29] Corinna Cortes and Vladimir Vapnik. 'Support-vector networks.' In: *Machine learning* 20.3 (1995), pp. 273–297.

[30] George Cybenko. 'Approximation by superpositions of a sigmoidal function.' In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[31] Ingrid Daubechies. *Ten lectures on wavelets*. Vol. 61. Siam, 1992.

[32] Marco De Curtis and Giuliano Avanzini. 'Interictal spikes in focal epileptogenesis.' In: *Progress in neurobiology* 63.5 (2001), pp. 541–567.

[33] Christine De Mol, Ernesto De Vito, and Lorenzo Rosasco. 'Elastic-net regularization in learning theory.' In: *Journal of Complexity* 25.2 (2009), pp. 201–230.

[34] Christine De Mol, Sofia Mosci, Magali Traskine, and Alessandro Verri. 'A regularized method for selecting nested groups of relevant genes from microarray data.' In: *Journal of Computational Biology* 16.5 (2009), pp. 677–690.

[35] Giancarlo Di Gennaro, Pier Paolo Quarato, Paolo Onorati, Giovanni B Colazza, Francesco Mari, Liliana G Grammaldo, Olga Ciccarelli, Nicolò G Meldolesi, Fabio Sebastiano, Mario Manfredi, et al. 'Localizing significance of temporal intermittent rhythmic delta activity (TIRDA) in drug-resistant focal epilepsy.' In: *Clinical Neurophysiology* 114.1 (2003), pp. 70–78.

[36] David L Donoho, Iain M Johnstone, et al. 'Minimax estimation via wavelet shrinkage.' In: *The annals of Statistics* 26.3 (1998), pp. 879–921.

[37] Mervyn J Eadie. 'Shortcomings in the current treatment of epilepsy.' In: *Expert review of neurotherapeutics* 12.12 (2012), pp. 1419–1427.

[38] Jerome Engel. 'Update on surgical treatment of the epilepsies: summary of the second international palm desert conference on the surgical treatment of the epilepsies (1992).' In: *Neurology* 43.8 (1993), pp. 1612–1612.

[39] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 'A density-based algorithm for discovering clusters in large spatial databases with noise.' In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231.

[40] Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. 'Regularization networks and support vector machines.' In: *Advances in computational mathematics* 13.1 (2000), p. 1.

[41] Tommaso Fedele, Sergey Burnos, Ece Boran, Niklaus Krayenbühl, Peter Hilfiker, Thomas Grunwald, and Johannes Sarnthein. 'Resection of high frequency oscillations predicts seizure outcome in the individual patient.' In: *Scientific reports* 7.1 (2017), p. 13836.

[42] Bruce Fischl. 'FreeSurfer.' In: *Neuroimage* 62.2 (2012), pp. 774–781.

[43] Jerome H Friedman. 'Greedy function approximation: a gradient boosting machine.' In: *Annals of statistics* (2001), pp. 1189–1232.

[44] Gene H Golub and Christian Reinsch. 'Singular value decomposition and least squares solutions.' In: *Linear Algebra*. Springer, 1971, pp. 134–151.

[45] Mehmet Gönen and Ethem Alpaydın. 'Multiple kernel learning algorithms.' In: *Journal of machine learning research* 12.Jul (2011), pp. 2211–2268.

[46] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[47] Robert Kent Goodrich. 'A Riesz representation theorem.' In: *Proceedings of the American Mathematical Society* 24.3 (1970), pp. 629–636.

[48] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen. 'MNE software for processing MEG and EEG data.' In: *Neuroimage* 86 (2014), pp. 446–460.

[49] Roger Grosse, Rajat Raina, Helen Kwong, and Andrew Y Ng. 'Shift-invariance sparse coding for audio classification.' In: *arXiv preprint arXiv:1206.5241* (2012).

[50] Isabelle Guyon and André Elisseeff. 'An introduction to variable and feature selection.' In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[51] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

[52] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. 'The elements of statistical learning: data mining, inference and prediction.' In: *The Mathematical Intelligencer* 27.2 (2005), pp. 83–85.

[53] Biyu J He, John M Zempel, Abraham Z Snyder, and Marcus E Raichle. 'The temporal structures and functional significance of scale-free brain activity.' In: *Neuron* 66.3 (2010), pp. 353–369.

[54] Alan L Hodgkin and Andrew F Huxley. 'A quantitative description of membrane current and its application to conduction and excitation in nerve.' In: *The Journal of physiology* 117.4 (1952), pp. 500–544.

[55] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. 'Kernel methods in machine learning.' In: *The annals of statistics* (2008), pp. 1171–1220.

[56] Yvonne Höller, Raoul Kutil, Lukas Klaffenböck, Aljoscha Thomschewski, Peter M Höller, Arne C Bathke, Julia Jacobs, Alexandra C Taylor, Raffaele Nardone, and Eugen Trinka. 'High-frequency oscillations in epilepsy and surgical outcome. A meta-analysis.' In: *Frontiers in human neuroscience* 9 (2015), p. 574.

[57] Aapo Hyvärinen and Erkki Oja. 'Independent component analysis: algorithms and applications.' In: *Neural networks* 13.4-5 (2000), pp. 411–430.

[58] Koji Iida and Hiroshi Otsubo. 'Stereoelectroencephalography: indication and efficacy.' In: *Neurologia medico-chirurgica* 57.8 (2017), pp. 375–385.

[59] Sergey Ioffe and Christian Szegedy. 'Batch normalization: Accelerating deep network training by reducing internal covariate shift.' In: *arXiv preprint arXiv:1502.03167* (2015).

[60] Julia Jacobs, Pierre LeVan, Rahul Chander, Jeffery Hall, François Dubeau, and Jean Gotman. 'Interictal high-frequency oscillations (80–500 Hz) are an indicator of seizure onset areas independent of spikes in the human epileptic brain.' In: *Epilepsia* 49.11 (2008), pp. 1893–1907.

[61] Julia Jacobs, Maeike Zijlmans, Rina Zelmann, Claude-Édouard Chatillon, Jeffrey Hall, André Olivier, François Dubeau, and Jean Gotman. 'High-frequency electroencephalographic oscillations correlate with outcome of epilepsy surgery.' In: *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 67.2 (2010), pp. 209–220.

[62] Julia Jacobs, Joyce Y Wu, Piero Perucca, Rina Zelmann, Malenka Mader, Francois Dubeau, Gary W Mathern, Andreas Schulze-Bonhage, and Jean Gotman. 'Removing high-frequency oscillations: a prospective multicenter study on seizure outcome.' In: *Neurology* 91.11 (2018), e1040–e1052.

[63] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. 'Data clustering: a review.' In: *ACM computing surveys (CSUR)* 31.3 (1999), pp. 264–323.

[64] Herbert H Jasper. 'Charting the sea of brain waves.' In: *Science* 108.2805 (1948), pp. 343–347.

[65] Dong Hyun Jeong, Caroline Ziemkiewicz, Brian Fisher, William Ribarsky, and Remco Chang. 'iPCA: An Interactive System for PCA-based Visual Analytics.' In: *Computer Graphics Forum*. Vol. 28. 3. Wiley Online Library. 2009, pp. 767–774.

[66] Jin Jing, Haoqi Sun, Jennifer A Kim, Aline Herlopian, Ioannis Karakis, Marcus Ng, Jonathan J Halford, Douglas Maus, Fonda Chan, Marjan Dolatshahi, et al. 'Development of Expert-Level Automated Detection of Epileptiform Discharges During Electroencephalogram Interpretation.' In: *JAMA neurology* 77.1 (2020), pp. 103–108.

[67] Philippa J Karoly, Dean R Freestone, Ray Boston, David B Grayden, David Himes, Kent Leyde, Udaya Seneviratne, Samuel Berkovic, Terence O?Brien, and Mark J Cook. 'Interictal spikes and epileptic seizures: their relationship and underlying rhythmicity.' In: *Brain* 139.4 (2016), pp. 1066–1078.

[68] Wlodzimierz Klonowski. 'Everything you wanted to ask about EEG but were afraid to get the right answer.' In: *Nonlinear biomedical physics* 3.1 (2009), p. 2.

[69] Bryan Kolb and Ian Q Whishaw. *Fundamentals of human neuropsychology*. Macmillan, 2009.

[70] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 'Imagenet classification with deep convolutional neural networks.' In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[71] Tom Dupré La Tour, Thomas Moreau, Mainak Jas, and Alexandre Gramfort. 'Multivariate convolutional sparse coding for electromagnetic brain signals.' In: *Advances in Neural Information Processing Systems*. 2018, pp. 3292–3302.

[72] Jean-Philippe Lachaux, Eugenio Rodriguez, Jacques Martinerie, and Francisco J Varela. 'Measuring phase synchrony in brain signals.' In: *Human brain mapping* 8.4 (1999), pp. 194–208.

[73] Gert RG Lanckriet, Tijl De Bie, Nello Cristianini, Michael I Jordan, and William Stafford Noble. 'A statistical framework for genomic data fusion.' In: *Bioinformatics* 20.16 (2004), pp. 2626–2635.

[74] Miroslaw Latka, Ziemowit Was, Andrzej Kozik, and Bruce J West. 'Wavelet analysis of epileptic spikes.' In: *Physical Review E* 67.5 (2003), p. 052902.

[75] Michel Le Van Quyen, Jack Foucher, Jean-Philippe Lachaux, Eugenio Rodriguez, Antoine Lutz, Jacques Martinerie, and Francisco J Varela. 'Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony.' In: *Journal of neuroscience methods* 111.2 (2001), pp. 83–98.

[76] Daniel D Lee and H Sebastian Seung. 'Learning the parts of objects by non-negative matrix factorization.' In: *Nature* 401.6755 (1999), p. 788.

[77] Gregory R Lee, Ralf Gommers, Filip Waselewski, Kai Wohlfahrt, and Aaron O?Leary. 'PyWavelets: A Python package for wavelet analysis.' In: *J. Open Source Softw* 4.36 (2019), p. 1237.

[78] Su Liu, Candan Gurses, Zhiyi Sha, Michael M Quach, Altay Sencer, Nerses Bebek, Daniel J Curry, Sujit Prabhu, Sudhakar Tummala, Thomas R Henry, et al. 'Stereotyped high-frequency oscillations discriminate seizure onset zones and critical functional cortex in focal epilepsy.' In: *Brain* 141.3 (2018), pp. 713–730.

[79] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[80] Stéphane Mallat. 'Understanding deep convolutional networks.' In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150203.

[81] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. 'Boosting algorithms as gradient descent.' In: *Advances in neural information processing systems*. 2000, pp. 512–518.

[82] H Matsumoto and C Ajmone Marsan. 'Cortical cellular phenomena in experimental epilepsy: ictal manifestations.' In: *Experimental neurology* 9.4 (1964), pp. 305–326.

[83] Manuel R Mercier, Stephan Bickel, Pierre Megevand, David M Groppe, Charles E Schroeder, Ashesh D Mehta, and Fred A Lado. 'Evaluation of cortical local field potential diffusion in stereotactic electroencephalography recordings: a glimpse on white matter signal.' In: *Neuroimage* 147 (2017), pp. 219–232.

[84] Sebastian Mika, Bernhard Schölkopf, Alex J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. 'Kernel PCA and de-noising in feature spaces.' In: *Advances in neural information processing systems*. 1999, pp. 536–542.

[85] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 'Efficient estimation of word representations in vector space.' In: *arXiv preprint arXiv:1301.3781* (2013).

[86] Partha P Mitra and Bijan Pesaran. 'Analysis of dynamic brain imaging data.' In: *Biophysical journal* 76.2 (1999), pp. 691–708.

[87] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 'Human-level control through deep reinforcement learning.' In: *Nature* 518.7540 (2015), pp. 529–533.

[88] Anne H Mooij, Birgit Frauscher, Mina Amiri, Willem M Otte, and Jean Gotman. 'Differentiating epileptic from non-epileptic high frequency intracerebral EEG signals with measures of wavelet entropy.' In: *Clinical Neurophysiology* 127.12 (2016), pp. 3529–3536.

[89] Massimo Narizzano, Gabriele Arnulfo, Serena Ricci, Benedetta Toselli, Martin Tisdall, Andrea Canessa, Marco Massimo Fato, and Francesco Cardinale. 'SEEG assistant: a 3DSlicer extension to support epilepsy surgery.' In: *BMC bioinformatics* 18.1 (2017), p. 124.

[90] Andrew Y Ng. 'Feature selection, L^1 vs. L^2 regularization, and rotational invariance.' In: *Proceedings of the twenty-first international conference on Machine learning*. ACM. 2004, p. 78.

[91] Johannes Niediek, Jan Boström, Christian E Elger, and Florian Mormann. 'Reliable analysis of single-unit recordings from the human brain under noisy conditions: tracking neurons over hours.' In: *PloS one* 11.12 (2016).

[92] Jeffrey Noebels, Massimo Avoli, Michael Rogawski, Richard Olsen, and Antonio Delgado-Escueta. *Jasper's basic mechanisms of the epilepsies*. Oxford University Press, 2012, p. 304.

[93] Antoine Nonclercq, Martine Foulon, Denis Verheulpen, Cathy De Cock, Marga Buzatu, Pierre Mathys, and Patrick Van Bogaert. 'Cluster-based spike detection algorithm adapts to interpatient and intrapatient variation in spike morphology.' In: *Journal of neuroscience methods* 210.2 (2012), pp. 259–265.

[94] Ibrahim Omerhodzic, Samir Avdakovic, Amir Nuhanovic, and Kemal Dizdarevic. 'Energy distribution of EEG signals: EEG signal wavelet-neural network classifier.' In: *arXiv preprint arXiv:1307.7897* (2013).

[95] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 'Wavenet: A generative model for raw audio.' In: *arXiv preprint arXiv:1609.03499* (2016).

[96] Neal Parikh, Stephen Boyd, et al. 'Proximal algorithms.' In: *Foundations and Trends® in Optimization* 1.3 (2014), pp. 127–239.

[97] Karl Pearson. 'LIII. On lines and planes of closest fit to systems of points in space.' In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.

[98] F. Pedregosa et al. 'Scikit-learn: Machine Learning in Python.' In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[99] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 'Scikit-learn: Machine learning in Python.' In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[100] Jyoti Pillai and Michael R Sperling. 'Interictal EEG and the diagnosis of epilepsy.' In: *Epilepsia* 47 (2006), pp. 14–22.

[101] John C Platt, Nello Cristianini, and John Shawe-Taylor. 'Large margin DAGs for multiclass classification.' In: *Advances in neural information processing systems*. 2000, pp. 547–553.

[102] Tomaso Poggio, Kenji Kawaguchi, Qianli Liao, Brando Miranda, Lorenzo Rosasco, Xavier Boix, Jack Hidary, and Hrushikesh Mhaskar. 'Theory of deep learning III: explaining the non-overfitting puzzle.' In: *arXiv preprint arXiv:1801.00173* (2017).

[103] Jan Pyrzowski, Mariusz Siemiński, Anna Sarnowska, Joanna Jedrzejczak, and Walenty M Nyka. 'Interval analysis of interictal EEG: pathology of the alpha rhythm in focal epilepsy.' In: *Scientific reports* 5 (2015), p. 16230.

[104] R Quian Quiroga, Zoltan Nadasdy, and Yoram Ben-Shaul. 'Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering.' In: *Neural computation* 16.8 (2004), pp. 1661–1687.

[105] Nicolas Roehri, Jean-Marc Lina, John C Mosher, Fabrice Bartolomei, and Christian-George Bénar. 'Time-frequency strategies for increasing high-frequency oscillation detectability in intracerebral EEG.' In: *IEEE Transactions on Biomedical Engineering* 63.12 (2016), pp. 2595–2606.

[106] Osvaldo A Rosso, Susana Blanco, Juliana Yordanova, Vasil Kolev, Alejandra Figliola, Martin Schürmann, and Erol Başar. 'Wavelet entropy: a new tool for analysis of short duration brain electrical signals.' In: *Journal of neuroscience methods* 105.1 (2001), pp. 65–75.

[107] Leonid I Rudin, Stanley Osher, and Emad Fatemi. 'Nonlinear total variation based noise removal algorithms.' In: *Physica D: nonlinear phenomena* 60.1-4 (1992), pp. 259–268.

[108] Andrew M Saxe, James L McClelland, and Surya Ganguli. 'Exact solutions to the nonlinear dynamics of learning in deep linear neural networks.' In: *arXiv preprint arXiv:1312.6120* (2013).

[109] Mona Sazgar and Michael G Young. 'EEG Artifacts.' In: *Absolute Epilepsy and EEG Rotation Review*. Springer, 2019, pp. 149–162.

[110] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. 'A generalized representer theorem.' In: *International conference on computational learning theory*. Springer. 2001, pp. 416–426.

[111] Gary Shaw and Dimitris Manolakis. 'Signal processing for hyperspectral image exploitation.' In: *IEEE Signal processing magazine* 19.1 (2002), pp. 12–16.

[112] SJM Smith. 'EEG in the diagnosis, classification, and management of patients with epilepsy.' In: *Journal of Neurology, Neurosurgery & Psychiatry* 76.suppl 2 (2005), pp. ii2–ii7.

[113] Alex J Smola and Bernhard Schölkopf. *Learning with kernels*. Vol. 4. Citeseer, 1998.

[114] KP Soman, R Loganathan, and V Ajay. *Machine learning with SVM and other kernel methods*. PHI Learning Pvt. Ltd., 2009.

[115] Perry Sprawls. *Magnetic resonance imaging: principles, methods, and techniques*. Medical Physics Publishing, 2000.

[116] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. 'Maximum-margin matrix factorization.' In: *Advances in neural information processing systems*. 2005, pp. 1329–1336.

[117] Nitish Srivastava. 'Improving neural networks with dropout.' In: *University of Toronto* 182.566 (2013), p. 7.

[118] Richard J Staba, Charles L Wilson, Anatol Bragin, Itzhak Fried, and Jerome Engel Jr. 'Quantitative analysis of high-frequency oscillations (80–500 Hz) recorded in human epileptic hippocampus and entorhinal cortex.' In: *Journal of neurophysiology* 88.4 (2002), pp. 1743–1752.

[119] Richard J Staba, Charles L Wilson, Anatol Bragin, Donald Jhung, Itzhak Fried, and Jerome Engel Jr. 'High-frequency oscillations recorded in human medial temporal lobe during sleep.' In: *Annals of neurology* 56.1 (2004), pp. 108–115.

[120] Richard J Staba, Leonardo Frighetto, Eric J Behnke, Gary W Mathern, Tony Fields, Anatol Bragin, Jennifer Ogren, Itzhak Fried, Charles L Wilson, and Jerome Engel Jr. 'Increased fast ripple to ripple ratios correlate with reduced hippocampal volumes and neuron loss in temporal lobe epilepsy patients.' In: *Epilepsia* 48.11 (2007), pp. 2130–2138.

[121] Kevin J Staley and F Edward Dudek. 'Interictal spikes and epileptogenesis.' In: *Epilepsy Currents* 6.6 (2006), pp. 199–202.

[122] Kevin J Staley, Andrew White, and F Edward Dudek. 'Interictal spikes: harbingers or causes of epilepsy?' In: *Neuroscience letters* 497.3 (2011), pp. 247–250.

[123] Jean Talairach. 'Approche nouvelle de la neurochirurgie de l'épilepsie: méthodologie stéréotaxique et résultats thérapeutiques.' In: *Congres An nuel de la Societe de Neurochirurgie de Langue Francaise. Marseille: Masson, 25-28 Juin*. 1974.

[124] Georgi P Tolstov. *Fourier series*. Courier Corporation, 2012.

[125] Veronica Tozzo, Vanessa D'Amario, and Annalisa Barla. 'Hey there's DALILA: a DictionAry LearnIng LibrAry.' In: *2017 Imperial College Computing Student Workshop (ICCSW 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2018.

[126] Joel A Tropp. 'Just relax: Convex programming methods for identifying sparse signals in noise.' In: *IEEE transactions on information theory* 52.3 (2006), pp. 1030–1051.

[127] Joel A Tropp and Anna C Gilbert. 'Signal recovery from random measurements via orthogonal matching pursuit.' In: *IEEE Transactions on information theory* 53.12 (2007), pp. 4655–4666.

[128] Eric Van Diessen, Judith I Hanemaaijer, Willem M Otte, Rina Zelmann, Julia Jacobs, Floor E Jansen, François Dubeau, Cornelis J Stam, Jean Gotman, and Maeike Zijlmans. 'Are high frequency oscillations associated with altered network topology in partial epilepsy?' In: *Neuroimage* 82 (2013), pp. 564–573.

[129] Rene Vidal, Joan Bruna, Raja Giryes, and Stefano Soatto. 'Mathematics of deep learning.' In: *arXiv preprint arXiv:1712.04741* (2017).

[130] Arthur A. Ward. 'The Epileptic Spike.' In: *Epilepsy* 1.1-5 (1959), pp. 600–606.

[131] Fabrice Wendling, Fabrice Bartolomei, Faten Mina, Clément Huneau, and Pascal Benquet. 'Interictal spikes, fast ripples and seizures in partial epilepsies–combining multi-level computational models with experimental data.' In: *European journal of Neuroscience* 36.2 (2012), pp. 2164–2177.

[132] Andreas Widmann, Erich Schröger, and Burkhard Maess. 'Digital filter design for electrophysiological data–a practical approach.' In: *Journal of neuroscience methods* 250 (2015), pp. 34–46.

[133] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 'On early stopping in gradient descent learning.' In: *Constructive Approximation* 26.2 (2007), pp. 289–315.

[134] Pierre Yger, Giulia LB Spampinato, Elric Esposito, Baptiste Lefebvre, Stéphane Deny, Christophe Gardella, Marcel Stimberg, Florian Jetter, Guenther Zeck, Serge Picaud, et al. 'A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo.' In: *Elife* 7 (2018), e34518.

[135] Rina Zelmann, F Mari, J Jacobs, M Zijlmans, R Chander, and J Gotman. 'Automatic detector of high frequency oscillations for human recordings with macroelectrodes.' In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE. 2010, pp. 2329–2333.

[136] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 'Understanding deep learning requires rethinking generalization.' In: *arXiv preprint arXiv:1611.03530* (2016).

[137] Maeike Zijlmans, Premysl Jiruska, Rina Zelmann, Frans SS Leijten, John GR Jefferys, and Jean Gotman. 'High-frequency oscillations as a new biomarker in epilepsy.' In: *Annals of neurology* 71.2 (2012), pp. 169–178.

[138] Hui Zou and Trevor Hastie. 'Regularization and variable selection via the elastic net.' In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.